EVALUATING THE VALIDITY AND COMPARABILITY OF PIRLS 2016 IN SOUTH AFRICA

Submitted to the University of Oxford
in partial fulfilment of the requirements
for the degree of Doctor of Philosophy in Education

Heather Leigh Kayton
Department of Education
Kellogg College

2024

Supervisors:
Dr Yasmine El Masri
Professor Victoria Murphy

"There are not many topics that encapsulate all of the different dimensions of South African inequality quite like that of reading."

(Spaull & Pretorius, 2019, p. 147)

Acknowledgements

This thesis would not have been possible without the support, guidance, and friendship of the community I am so fortunate to be a part of. I would like to thank everyone who has been on this journey with me.

Firstly, thank you to my master's supervisors at the University of Johannesburg, Prof Anne-Marie Beukes and Prof William Botha, for inspiring me to continue learning, researching, and to look beyond what I thought possible. I am also grateful to the First Rand International Postgraduate Scholarships Fund, the Oppenheimer Memorial Trust, and Kellogg College their financial support, which helped make the pursuit of this PhD possible.

To my supervisors at the Department of Education, Yasmine El Masri and Victoria Murphy, thank you for your invaluable guidance, advice, and generous feedback throughout this process. I have learned so much under your guidance. I will forever be grateful for the support you have given me and the trust you have had in me – particularly when I couldn't find it in myself.

To my colleagues at the Oxford University Center for Educational Assessment, thank you for giving me the opportunity to be part of projects that introduced me to the wider world and impact of educational assessments. And to my colleagues at Hertford College, thank you for giving me a home away from home, and for patiently listening to me complain about the weather all year round. I would also like to acknowledge, Jessica Briggs Baffoe-Djan, Robert Woore, Joshua McGrane, Michelle Meadows, Anil Kanjee, and Montserrat Gomendio for their careful review, comments, and constructive suggestions on earlier drafts of this thesis.

To my friend and mentor, Josh, thank you for making psychometrics feel attainable, for the faith you have in me, and for the time, support, and expertise you have so generously offered me. You are the voice in my head pushing me to do more and worry less. Thank you for making Oxford feel like home when everything felt isolating and unfamiliar.

To all of my friends in Oxford, thank you for inspiring me, you make everything feel achievable, even things I could never have presumed possible. I feel so fortunate to have found a community with you all, despite the world being closed for half of our time here. Your presence and support have been invaluable throughout this journey.

To my friends at home, thank you for always taking the time to make me feel like I am not so far away. Vuyelwa, thank you for keeping me grounded, and for showing me what it truly means to care about education in South Africa. Camilla and Gugu, thank you for always being willing to hop on a call and share your time, wisdom, and stories with me. Karen, thank you for your insightful advice, compassion, and encouragement whenever I reached out to you. Claudia, being away from you over these past few years has been so difficult, thank you for always being there for me no matter how difficult things were for you.

To my partner, Richard, thank you for your support, help, and love throughout this journey. I am so grateful that Oxford brought us together. Thank you for always saying yes when I asked you to read 'just this little section here', for being honest when it didn't make sense, and for reassuring and encouraging me when it did. I am not sure how (or if) I would have been able to make it through this without your patience, selflessness, generosity, and understanding.

To my family, moving away from you has been the most difficult thing I have ever done, thank you for always prioritising me no matter how far away I was. Mom and dad, thank you for sacrificing so much of your financial freedom to enable me to continue here when I thought it was the end of my journey. I will always be thankful for your unconditional support and encouragement. Daegan and Cami, thank you for calling me every morning even though I didn't always answer, you brighten up my days (which are very grey here). John, thank you for always making me feel like a part of your home, even when I was stuck thousands of miles away.

Finally, to Cheryl, thank you for being the most consistent, supportive, and dependable person in my life. I feel privileged to have you as my sister. Thank you for always taking my calls, listening to my rants, and for all the long rambling voice notes that make me feel like I am in the car with you. Also, an extra thank you, Cheryl, for instilling in me (many years ago) what rigorous and ethical research looks like.

Contents

CHAPTER 1 INTRODUCTION **10**

1.1 OVERVIEW OF PIRLS 2016 13
1.2 PIRLS 2016 RESULTS IN SOUTH AFRICA 14
1.3 OUTLINE OF THIS INTEGRATED THESIS 17

CHAPTER 2 CONTEXTUAL CONSIDERATIONS **22**

2.1 APARTHEID EDUCATION POLICIES 22
2.2 CURRENT EDUCATION POLICIES, SYSTEMS AND PRACTICES 24
2.2.1 Language in education policies *25*
2.3 LARGE-SCALE READING ASSESSMENT IN SOUTH AFRICA 28
2.3.1 Cross-national assessment studies *28*
2.3.2 National assessment studies *29*
2.3.3 Provincial assessment studies *31*
2.4 SOUTH AFRICA'S PARTICIPATION IN PIRLS 33
2.4.1 PIRLS reading achievement in South Africa *33*
2.4.2 Status of PIRLS in South Africa *37*
2.4.3 Situating PIRLS results in the context of South African schools *40*

CHAPTER 3 LITERATURE REVIEW **42**

3.1 CONSIDERING THE CONSTRUCT 42
3.1.1 Theoretical frameworks of reading comprehension *43*
3.1.2 Challenges with assessing reading comprehension across languages *45*
3.1.3 The PIRLS reading framework *47*
3.2 CONSIDERING THE COMPONENTS OF QUALITY ASSESSMENT 48
3.2.1 Conceptualising validity *49*
3.2.2 Conceptualising reliability *51*
3.2.3 Conceptualising fairness *53*
3.3 ENSURING QUALITY FOR INTERNATIONAL LARGE-SCALE ASSESSMENTS 58
3.4 KEY AREAS OF CONCERN FOR PIRLS IN SOUTH AFRICA 60
3.4.1 Threat 1 – Contextual relevance *61*
3.4.2 Threat 2 – Wide gaps in achievement *63*
3.4.3 Threat 3 – Comparability across language versions *65*
3.4.4 Summary *67*

CHAPTER 4 METHODOLOGY **68**

4.1 DATA 68
4.1.1 Sample *69*
4.1.2 Instrument data *73*
4.2 EXPLORING READING ACHIEVEMENT IN THE SOUTH AFRICAN CONTEXT 73
4.3 EVALUATING THE PSYCHOMETRIC PROPERTIES OF PIRLS LITERACY 2016 IN SOUTH AFRICA 74
4.3.1 Selecting language groups *74*
4.3.2 Item response theory framework *75*
4.3.3 Evaluating fit *77*
4.3.4 Evaluating comparability *78*
4.4 METHODOLOGICAL LIMITATIONS AND CONSIDERATIONS 81
4.4.1 Questionnaire data *82*
4.4.2 Methodological approaches *83*

CHAPTER 5 (PAPER 1) LANGUAGE MATTERS: PIRLS READING ACHIEVEMENT IN SOUTH AFRICA 85

5.1 INTRODUCTION 85
5.1.1 PIRLS in South Africa 87
5.1.2 Aims and research questions 91
5.2 LITERATURE REVIEW 92
5.2.1 Reading comprehension 92
5.2.2 Assessing reading comprehension 93
5.2.3 Factors influencing reading comprehension 94
5.2.4 Limitations of current evidence base 98
5.3 METHODS 99
5.3.1 Data 99
5.3.2 Multiple linear regression 101
5.4 ANALYSIS 105
5.5 RESULTS 107
5.5.1 Descriptive statistics 107
5.5.2 Regression results 110
5.5.3 Modelling contextual factors 111
5.5.4 Modelling language factors 115
5.6 DISCUSSION 124
5.7 LIMITATIONS AND FUTURE RESEARCH 129

CHAPTER 6 (PAPER 2) FIT FOR PURPOSE OR OFF THE MARK: PIRLS 2016 IN SOUTH AFRICA .. 132

6.1 INTRODUCTION 133
6.1.1 What is PIRLS? 134
6.1.2 PIRLS 2016 in South Africa 135
6.1.3 The purpose of participation in international large-scale assessments 136
6.1.4 Aim and research questions 138
6.2 LITERATURE REVIEW 138
6.2.1 Widening participation in international large-scale assessments 138
6.2.2 Fundamentals of quality in large-scale assessments 140
6.3 METHODOLOGY 144
6.3.1 Data 144
6.3.2 Item response theory models 145
6.4 ANALYSIS 147
6.4.1 Evaluating model fit 148
6.4.2 Checking model assumptions 148
6.4.3 Examining targeting 149
6.5 RESULTS 150
6.5.1 Model fit 150
6.5.2 Dimensionality 151
6.5.3 Dependence 153
6.5.4 Targeting 154
6.6 DISCUSSION 162
6.7 LIMITATIONS AND FUTURE RESEARCH 166

CHAPTER 7 (PAPER 3) DIF HAPPENS: THE IMPACT OF DIFFERENTIAL ITEM FUNCTIONING ACROSS LANGUAGE VERSIONS OF PIRLS 2016 IN SOUTH AFRICA 168

7.1 INTRODUCTION 168

7.1.1 Aim and research questions ... *170*
7.2 Literature review ... 171
7.2.1 Ensuring quality in international large-scale assessments ... *171*
7.2.2 Cross-language comparability ... *172*
7.2.3 The overall impact of differential item functioning ... *174*
7.3 Methods ... 175
7.3.1 Data ... *175*
7.3.2 Passages and items in PIRLS Literacy 2016 ... *176*
7.3.3 Detecting differential item functioning ... *178*
7.3.4 Evaluating the impact of differential item functioning ... *180*
7.4 Analysis ... 182
7.4.1 Item fit evaluation ... *182*
7.4.2 Differential item functioning analysis ... *183*
7.4.3 Differential item functioning impact analysis ... *184*
7.5 Results ... 184
7.5.1 Item response theory model estimation and fit evaluation ... *184*
7.5.2 Determining comparability at an item level ... *186*
7.5.3 Evaluating differential item functioning impact at a passage level ... *188*
7.5.4 Evaluating differential item functioning impact at test level ... *196*
7.6 Discussion ... 198
7.7 Limitations and future research ... 203

CHAPTER 8 DISCUSSION ... 206

8.1 Missing the mark ... 207
8.1.1 Establishing validity ... *207*
8.1.2 Evaluating reliability ... *212*
8.1.3 Considering fairness ... *214*
8.2 Mapping the future ... 216
8.2.1 Lessons learned ... *216*
8.2.2 Future prospects ... *220*

CHAPTER 9 REFLECTIONS ... 226

9.1 Limitations ... 226
9.2 Contributions ... 228
9.3 Implications ... 231
9.3.1 Implications for international large-scale assessment development and design ... *231*
9.3.2 Implications for the use of international large-scale assessment results ... *233*

REFERENCES ... 239

APPENDICES ... 292

Appendix A PIRLS 2016 instrument ... 292
A-1 Description of PIRLS Literacy 2016 passages and items ... 292
A-2 Booklet Design ... 295
A-3 Scoring ... 296
A-4 PIRLS Benchmarks ... 300
Appendix B Data preparation and recoding ... 302
Appendix C Paper 1 additional information ... 306
C-1 Scaled items ... 306
C-2 Assumptions and diagnostics of the regression model ... 312
C-3 Detailed result tables for regression Models 0 to 5 ... 320

APPENDIX D PAPER 2 ADDITIONAL INFORMATION 334
D-1 ITEM RESPONSE FUNCTIONS FOR MISFITTING ITEMS 334
D-2 ITEM RESPONSE FUNCTIONS FOR NEGATIVELY DISCRIMINATING ITEMS 337
APPENDIX E PAPER 3 ADDITIONAL INFORMATION 349
E-1 PASSAGE-LEVEL DIFFERENTIAL RESPONSE FUNCTIONING STATISTICS 349

Tables

Table 2-1 *Overview of Grade 4 PIRLS results in South Africa 2006-2021* 36
Table 5-1 *Description of predictor variables and their source questionnaires* 104
Table 5-2 *Descriptive statistics for regression analysis variables* 107
Table 5-3 *Summary of multiple regression model results* 111
Table 5-4 *Results of multiple regression models comparing coefficients and t-statistics* 119
Table 6-1 *Comparison of fit statistics for simple and mixed IRT models* 150
Table 6-2 *Comparison of fit statistics for uni- and multidimensional IRT models* 151
Table 6-3 *Misfitting items in both uni- and multidimensional models* 152
Table 6-4 *Items related to secondary factors in multidimensional model* 153
Table 6-5 *Item pairs flagged for local dependence* 154
Table 6-6 *Count of poorly discriminating items by group* 158
Table 7-1 *Summary of passages and items in PIRLS Literacy 2016* 178
Table 7-2 *Model fit statistics before and after item set reduction* 185
Table 7-3 *Summary of effect size and direction of differential item functioning by focal group* 187
Table 7-4 *Passage-level differential response functioning statistics by focal group* 191
Table 7-5 *Test-level differential response functioning statistics by focal group* 196
Table A-1 *PIRLS Literacy 2016 passage titles and description* 293
Table A-2 *Matrix design for passage distribution across booklets* 296
Table A-3 *Item types, scoring procedures and models applied in PIRLS Literacy 2016* 297
Table A-4 *Description of benchmark criteria for PIRLS 2016* 300
Table A-5 *Item response original response and recoded values* 302
Table A-6 *Occurrences of recoded missing data by language group* 305
Table A-7 *Summary of item sets for scaled items and scale reliability statistics* 306
Table A-8 *Variance inflation factors for variables in Model 5* 316
Table A-9 *Results of multiple linear regression Model 0* 320
Table A-10 *Results of multiple linear regression Model 1* 321
Table A-11 *Results of multiple linear regression Model 2* 322
Table A-12 *Results of multiple linear regression Model 3* 324
Table A-13 *Results of multiple linear regression Model 4* 327
Table A-14 *Results of multiple linear regression Model 5* 330
Table A-15 *Item parameters for freely estimated model* 340
Table A-16 *Differential response functioning statistics at the passage level – Afrikaans focal group* 349
Table A-17 *Differential response functioning statistics at the passage level – Sepedi focal group* 350

Figures

Figure 1-1 Reading achievement levels across language groups in PIRLS Literacy 2016 16
Figure 4-1 Number of students per language group in the PIRLS 2016 South African dataset 70
Figure 4-2 Frequency of test language use at home by South African students 72
Figure 5-1 PIRLS 2016 international benchmark achievement by language group in South Africa 89
Figure 5-2 Distribution of students per language group in the PIRLS 2016 South African dataset 100
Figure 6-1 Percentage of students responding correctly to items in each group 155
Figure 6-2 Expected versus observed item response functions – examples of good and poor fit 157
Figure 6-3 Expected versus observed item response functions – examples of negatively discriminating items across groups 158
Figure 6-4 Wright Map showing distribution of student ability and distribution of item difficulty ... 160
Figure 6-5 Annotated Wright Map highlighting misaligned items for all three groups...................... 161
Figure 7-1 Histogram of average PIRLS Literacy 2016 scores for each language group 176
Figure 7-2 Items with large differential item functioning – Afrikaans focal group 188
Figure 7-3 Examples of items with large differential item functioning – Sepedi focal group 188
Figure 7-4 Summary of item functioning and fit per passage 190
Figure 7-5 Differential response functioning plots for Sepedi focal group – Passages G, H, J and K . 194
Figure 7-6 Differential response functioning plots for Afrikaans focal group – Passages B, E and G . 195
Figure 7-7 Expected test score functions and sDRF plots for both focal groups (n = 152 items) 197
Figure A-1 Example excerpt and items from "The Pearl" by Mary Joslin 294
Figure A-2 Example of constructed response items from "The Pearl" 298
Figure A-3 Example of a scoring guide for constructed response items for "The Pearl" 299
Figure A-4 Partial-residual plots for continuous predictor variables 313
Figure A-5 Residuals of regression model versus fitted values for Model 5 314
Figure A-6 Correlation matrix for predictor variables in Model 5 315
Figure A-7 Scale-location plot of standardised residuals for Model 5 318
Figure A-8 Histogram of model residuals for Model 5 318
Figure A-9 Normal QQ-plot of model residuals for Model 5 319
Figure A-10 Expected versus observed item response functions for 21 misfitting items 336
Figure A-11 Expected versus observed item response functions for negatively discriminating items across groups 339

Equations

Equation 1: 2-Parameter Logistic IRT Model 146
Equation 2: Generalised Partial Credit IRT Model 146
Equation 3: Signed differential response function 181
Equation 4: Unsigned differential response function 181

Abstract

Large-scale assessments of reading, such as the Progress in International Reading Literacy Study (PIRLS), are useful for understanding reading development. PIRLS aims to provide valid and reliable assessments of reading across multiple countries. However, the extreme range in reading achievement outcomes, from the very high performance of some countries to the very low performance of others, creates substantial disparities across countries that threaten the validity, reliability, and fairness of PIRLS results. South Africa participates in PIRLS to better understand and inform interventions that can help improve reading in the country. But consistently low performance in PIRLS, coupled with intricate multilingualism, historical racial discrimination, and extreme inequality present further challenges to the already complicated task of assessing reading.

This thesis evaluates the validity and comparability of PIRLS 2016 in the unique local context of South Africa. It considers the implications of factors such as low overall performance, performance differences across groups, fit of the analytical models used, and comparability of item difficulty across language versions for the validity of the test and items. Because of the complexities associated with large-scale assessment in diverse contexts, the thesis brings together three papers, each focused on a specific area of concern regarding PIRLS 2016 in South Africa. While highlighting some concerns, the papers also discuss the implications that factors such as low performance and language diversity have for the valid interpretation of PIRLS scores, as well as the implications for the use of PIRLS to inform education policy and practice in South Africa.

The first paper examines the relationship between test language and reading achievement, revealing that test language significantly influences overall reading achievement in PIRLS 2016 in South Africa, even when controlling for other influential contextual factors. The second paper evaluates the validity of the PIRLS 2016 test instrument for South African students and highlights substantial issues with the alignment between item difficulty and student ability. The third paper investigates comparability across language versions and uncovers substantial evidence of comparability concerns at the item and passage level.

Overall, the findings from this thesis raise concerns about the use of PIRLS data to inform educational policy in South Africa. By highlighting key challenges in the South African context, this thesis contributes to a deeper understanding of the importance of addressing the needs of unique local contexts to ensure large-scale assessments are able to provide valid, reliable, and fair information for all participating countries.

CHAPTER 1 INTRODUCTION

The value of reading proficiency cannot be overstated. Reading is not only a skill that lays the foundation for educational success; it is a gateway to further opportunities and serves as a fundamental tool to enable individuals to fully participate in society (Aukerman & Chambers Schuldt, 2021). Understanding how best to develop reading proficiency is therefore important for all students. This is particularly true in low-income countries, where access to quality education is sometimes limited and reading literacy levels remain a challenge (Chmielewski, 2019; Hanushek et al., 2019). To gain insights into the effectiveness of educational systems, as well as to develop and improve policies to support successful reading development, many countries turn to international large-scale assessments (ILSA) of reading such as the Organization for Economic Cooperation and Development's (OECD) Programme for International Student Achievement (PISA) and the International Association for the Evaluation of Educational Achievement's (IEA) Progress in International Reading Literacy Study (PIRLS). Reading assessments can offer valuable insights into the nature and development of reading in context, as well as guidance for enhancing reading practice (Rojas-Torres et al., 2021; Schildkamp, 2019).

However, effectively assessing a student's reading skills is not a straightforward task (Araújo et al., 2013; Geva et al., 2019). Reading is a complex skill that encompasses a number of linguistic, cognitive, and sociocultural factors and processes. This means that assessing it necessitates a comprehensive, multifaceted approach. PIRLS is recognised as one of the world's leading reading assessment studies (Lenkeit et al., 2015; Stiff et al., 2023). PIRLS adopts a robust assessment framework that not only acknowledges the intricate, multifaceted nature of reading but also incorporates comprehensive questionnaires to capture broader social and contextual influences (Mullis et al., 2015a). Moreover, its implementation across diverse countries and contexts, with dedicated questionnaires on curriculum and policy, allows for informed monitoring, comparison, and evaluation of educational policies across borders (Hooper et al., 2015).

Assessing any area of academic performance across different countries introduces a number of challenges. Quality assessment of academic performance needs to ensure that it can accurately and effectively measure the intended educational outcomes at the required scale. However, performance in PIRLS varies considerably across countries, with some countries achieving remarkably high outcomes and others demonstrating extremely low outcomes (Mullis et al., 2012, 2017a, 2023). These extensive disparities in performance pose challenges to the validity and reliability of the assessment instruments as well as the analytic procedures used to draw insights from the assessment data. Researchers have shown that, particularly for low-performing countries, the estimation procedures used to calculate results can be adversely affected by wide ranges in achievement across participating countries (Rutkowski et al., 2019; Tijmstra et al., 2020; Wagner et al., 2011). It is crucial to recognise that these differences, as well as the heterogeneity within contexts, can impact the interpretation and generalisability of PIRLS results. These differences should not be overlooked, particularly in contexts where ILSAs are used as a basis for monitoring and evaluating reading achievement at a national level and are used to inform policy in some contexts (Gomendio, 2021; Lockheed & Wagemaker, 2013; Mulongo & Amod, 2019; van der Berg et al., 2016).

Throughout its participation in four PIRLS cycles, South Africa has consistently had alarmingly low performance; indicative of a persistent and deeply concerning issue often referred to as "the reading crisis" (Andersen, 2017; Howie, Combrinck, Roux, Mokoena, et al., 2017; Janks, 2011; Spaull & Pretorius, 2019). However, this crisis extends beyond reading proficiency alone, it encompasses a broader range of challenges within the South African educational system; including low performance across different subject areas, inadequate infrastructural facilities, limitations in teaching quality and capacity, as well as high attrition rates within the education sector (Hart & Zinn, 2015; Motala & Carel, 2019; Naidoo et al., 2014; Spaull & Jansen, 2019; Taylor, 2019; van der Berg et al., 2011). These difficulties are rooted in the historical legacy of racial discrimination that characterised apartheid-era education policies, highlighting the ongoing struggle of the education system in South Africa as it endeavours to recover from the inequities of the past (Fiske & Ladd, 2004; Gustafsson & Taylor, 2022; Kanjee & Sayed, 2013; Salisbury, 2016; Taylor & Yu, 2009; van der Berg & Gustafsson, 2019). Although navigating through these challenges is undeniably complex, it is crucial to recognise the profound influence of reading as a fundamental skill that underpins educational outcomes across diverse domains (Chapman et al., 2000). Consequently, the question of “what can be done” to improve reading outcomes becomes central, but answering this question requires rigorous, valid, reliable, and fair data.

In addition to the challenges South Africa faces with vastly unequal educational outcomes resulting from apartheid era discrimination policies; the complexity of the South African context is further compounded by its rich, diverse multilingualism. With 11 official languages[1], South Africa is a linguistically intricate society and students have the right to education in any of the official languages, which means 11 languages serve as media of instruction in the primary school. Under the apartheid regime, language in education policies were employed as a tool for segregation, discrimination, and oppression; new language in education policies now aim to serve as a means of redress and justice for previously marginalised language groups (Department of Basic Education (DBE), 2013; Heugh, 2013; Mutekwe & Sedibe, 2015; Nkadimeng & Makalela, 2023). This linguistic landscape adds an additional layer of complexity to the task of assessing reading effectively and comparably across languages in South Africa.

PIRLS is currently an important tool in South Africa as it is one of the only large-scale assessments of reading conducted at a national scale in 11 official languages. As such, it has become one of the primary instruments for monitoring the progress of reading in early grades (DBE, 2023). However, given the fundamental role of language in reading, navigating the complexities of multilingualism in South Africa is imperative for ensuring assessments can capture reading

[1] In May 2023, an amendment to the Constitution of South Africa was approved to include South African Sign Language as a 12th official language, however this has not yet been signed into law (Constitution Review Committee, 2017; Parliament of the Republic of South Africa, 2023).

achievement across different languages fairly and reliably. In the case of PIRLS, the test is routinely and rigorously evaluated by developers, as well as researchers conducting secondary analyses, who typically compare differences across countries, however, the same attention is not often given to differences across subgroups within countries (e.g., Elosua Oliden & Mujika Lizaso, 2013; Ercikan & Por, 2020; Finch & Hernández Finch, 2016; Goodrich & Ercikan, 2019; Roux et al., 2022). The diverse linguistic and cultural differences between students in South Africa warrant deeper investigation into these systematic differences within the country.

1.1 Overview of PIRLS 2016

PIRLS is designed to assess the reading ability of students in their fourth year of schooling (usually around 10 years old) across multiple countries and contexts, where reading ability is considered a purpose-driven, constructive, and interactive process (Mullis et al., 2015b). PIRLS is designed and run by the IEA and has been implemented internationally every five years since 2001. In 2016, PIRLS was conducted in 50 countries and approximately 319,000 students participated in the study (Mullis et al., 2017a). The reading test consists of 12 reading passages (each between 400 – 800 words), each with approximately 15 associated items that aim to address two reading purposes; reading for literacy experience and reading to acquire and use information (Mullis et al., 2015b).

There are three distinct tests: PIRLS 2016 (the main international study), ePIRLS 2016 (an assessment of online reading), and PIRLS Literacy 2016. PIRLS Literacy 2016 is designed to be a less difficult, but still scalable, version of the PIRLS assessment. It aims to provide further insight into the reading achievement of students at the lower end of the PIRLS achievement scale by presenting shorter passages with less complex syntax, interspacing the items within the passages to allow students to respond as they are reading, as well as adjusting the format of the passages to have a larger font and the style of a text aimed at younger readers (Martin, Mullis, & Foy, 2017). This study focuses on PIRLS Literacy 2016 as it was used as the primary instrument to assess Grade 4 students in South Africa, and as it was the only version conducted in 11 official languages (Howie, Combrinck, Roux, Mokoena, et al., 2017).

Each student receives one booklet, which contains two reading passages and approximately 30 items (~15 per passage). Both multiple choice and constructed response items are included, and items are scored either dichotomously (0 – incorrect, 1 – correct) or polytomously (e.g., 0 – incorrect, 1 – partially correct, 2 – correct). A comprehensive scoring guide and training sessions are provided by the IEA to support scoring consistency for the constructed response items. The reliability of scoring within countries is checked internationally; a subset of 200 student responses is randomly selected by each participating country to be scored independently by two scorers and the degree of agreement between these independent scorers is then reported as a measure of reliability. If the inter-rater reliability is below 85%, scorers are required to retrain and then rescore items. The average inter-rater reliability across items for PIRLS Literacy 2016 in South Africa was 93% (Martin, Mullis, & Hooper, 2017).

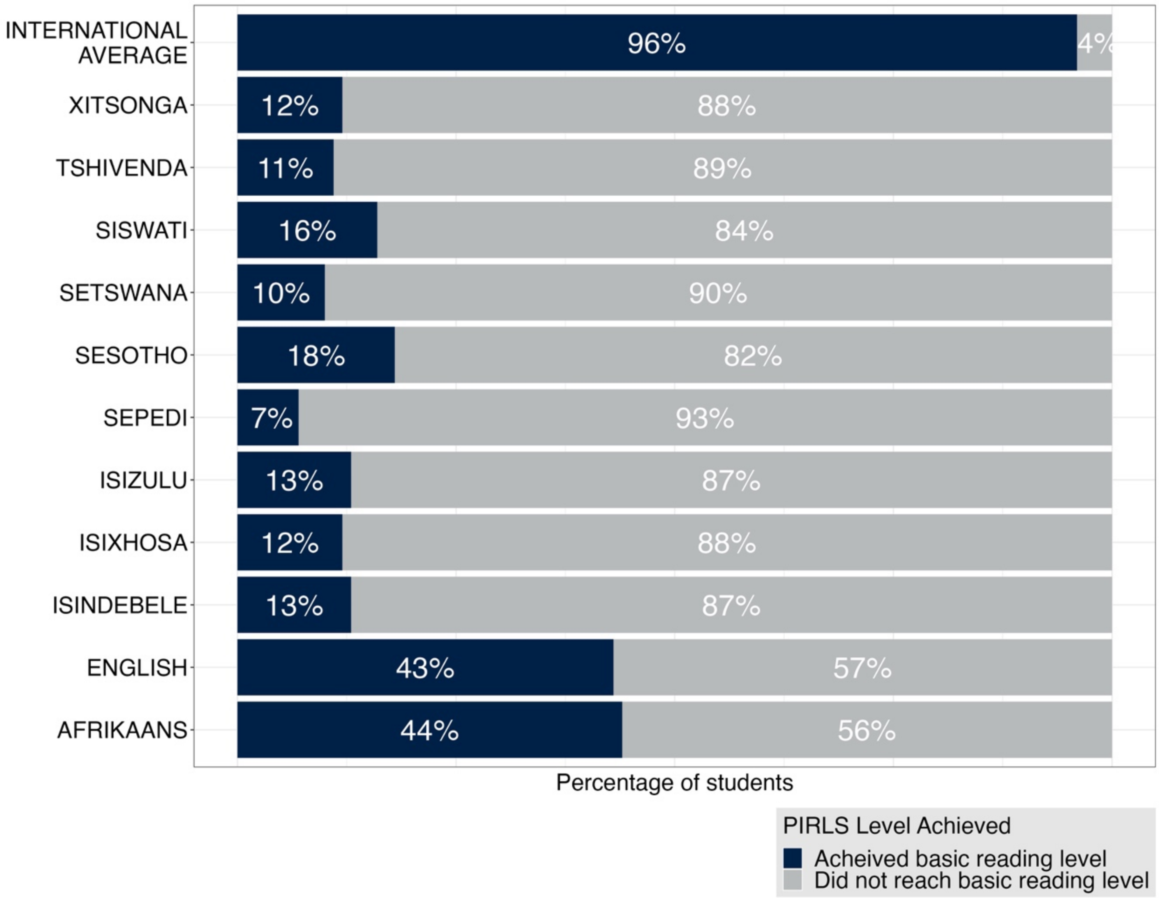
Following scoring and processing of the results, PIRLS provides an overall score for reading achievement for each student along a linear scale with a mean of 500 points and a standard deviation of 100 points. A more detailed description of the PIRLS 2016 instrument as well as examples of passages and items are available in Appendix A.

1.2 PIRLS 2016 results in South Africa

To assist with the interpretation of the scores, PIRLS sets out four benchmarks of reading achievement each associated with different levels of reading ability: low, intermediate, high, and advanced. Each of these benchmarks corresponds with a set of criteria that describes what students at each benchmark are able to demonstrate when reading literary and informational texts. Students who do not reach the low benchmark are not able to retrieve basic information from the text and consequently are deemed unable to read for meaning (see Appendix A-4 for further details on PIRLS benchmark criteria).

In 2016, Grade 4 students from South Africa participated in PIRLS Literacy and achieved an average score of 320, well below the international centerpoint of 500, with 78% of students scoring below the low benchmark level (Mullis et al., 2017a). This means that only 22% of students were able to demonstrate the ability to read at a basic level after four years of formal schooling (Howie, Combrinck, Roux, Mokoena, et al., 2017). This proportion is substantially lower than the international average; where across all 50 participating countries just 4% of students scored below the low benchmark level. These results were discussed widely in the media and research circles, and considerable concern was raised about the state of education in South Africa which had already been generally regarded as a system in crisis (Andersen, 2017; DBE, 2017a; Sindi-Leigh, 2019; Spaull, 2019; Taylor, 2017). However, because of the stark inequalities in the South African context, considering average performance is often not very meaningful, and sometimes quite misleading (Reddy, 2010; Spaull & Pretorius, 2019). As a result, the focus of concern regarding the South African results centred around the inequalities that are evident when looking at differences in reading performance across groups within the country. The language in which students wrote the test and wealth inequalities across fee-paying and no-fee schools were the primary factors seen to influence reading performance in PIRLS Literacy 2016 (Howie, Combrinck, Roux, Mokoena, et al., 2017). Similar patterns of inequalities were seen in South African reading achievement results when comparing results by school wealth, and results by test language (Spaull & Pretorius, 2019).

In 2016 in South Africa, students wrote the test in the language that was the medium of instruction at their school for the first three years, which is typically the students’ home language (usually the dominant language of the geographical area) for the first three years of primary school. This means that a majority (>70%) of South African students begin to learn to read in an African language (Spaull, 2022). Figure 1-1 shows the within country variation in reading achievement across language groups for PIRLS Literacy 2016 in South Africa. Almost half of the students who took the test in English or Afrikaans (43-44%) reached the low benchmark, while less than one fifth of students who took the test in one of the nine African languages (7-18%) reached the same benchmark. Students who took the test in Sepedi (Sesotho sa Leboa)[2] were the lowest performing group, with only 7% of students demonstrating the ability to read at a basic level.



Source: IEA International Data Repository for PIRLS 2016

Figure 1-1 Reading achievement levels across language groups in PIRLS Literacy 2016

These differences are not unexpected given the linguistic landscape in South African schools is complex and still recovering from decades of targeted racial discrimination that severely disadvantaged speakers of African languages. Since most South Africans are home language speakers of an African language and attend schools that use an African language as the medium of instruction for the first three years; understanding these differences is an important step in redressing the historical discrimination that continues to disproportionately affect students.

[2] There is contention surrounding the inconsistent terminology used for the Sepedi/Sesotho sa Leboa official languages in the interim (1993) and final (1996) Constitution documents. Some feel that Sepedi was erroneously declared as an official language as it is a variety of Sesotho sa Leboa (Northern Sotho) (Constitution Review Committee, 2017). This thesis, in line with the Constitution of the Republic of South Africa 1996, refers to Sepedi but recognises that this remains an unresolved issue (Rakgogo & Zungu, 2022).

However, if important decisions and policy changes are going to be made based on the results of a test, interrogating the applicability of the test for the students and context is essential to ensure that the results being relied upon for decision making are indeed valid and comparable. Despite this need, there has been limited critical investigations of ILSAs like PIRLS for unique multilingual contexts like South Africa. This thesis aims to fill this gap by thoroughly evaluating key concerns regarding the validity and comparability of information from PIRLS 2016 data, particularly as it is used to inform educational policy and improve learning outcomes in the complex South African context.

1.3 Outline of this integrated thesis

This research investigates the PIRLS Literacy 2016 data for South Africa, with the goal of extracting valuable insights, assessing the validity and comparability of the information provided, and identifying any potential issues with using PIRLS in the South African context. The overarching aim of this research is to answer the question of what can be learned from PIRLS Literacy 2016 about the reading achievement of South African students across different languages. The insights gained from this research will contribute to the development of improved assessment measures and facilitate more robust, valid, and fair evaluations of reading in South Africa. To evaluate the broad landscape of reading assessment in South Africa critically and comprehensively, this thesis combines three distinct papers, each dedicated to addressing distinct research questions. The information below briefly describes the contents and scope of each chapter in this integrated thesis.

Chapter 1 is the introduction to the thesis. It establishes the importance of reading assessments and presents a description of PIRLS, including how it is administered and scored. It describes the very low and unequal PIRLS Literacy 2016 results for South Africa, which motivate the research questions under investigation in each paper. This chapter also presents the overarching research aims and establishes the rationale for investigating PIRLS data for use in the South African context.

Chapter 2 presents an overview of relevant contextual considerations specific to educational achievement in South Africa. In particular, it describes the historical and current language in education policies and their continued impact on schooling and educational outcomes in the country. Also included in this chapter is an overview of large-scale reading assessments in South Africa as well as details of South Africa's participation in PIRLS.

Chapter 3 presents a review of the literature related to fundamental concepts relevant to the overall thesis. It begins by examining theoretical frameworks of reading comprehension and the challenges of assessing reading across languages, including an overview of the PIRLS reading framework. The chapter then provides an overview of some of the key literature related to the essential components of quality assessment: validity, reliability, and fairness, as well as approaches for ensuring quality in ILSAs. The chapter concludes with a discussion of some of the primary threats to the validity and comparability of PIRLS in the South African context that have motivated the investigations undertaken in each paper. As this is an integrated thesis, the primary literature relevant to each paper is included in the papers themselves (presented as Chapters 4, 5 & 6).

Chapter 4 provides an overview of the methodological approach taken to evaluate PIRLS in South Africa. This chapter also provides details about the data used for the thesis as a whole as well as the subsets of data considered in each paper. A full description of the methodology and analytic techniques used for each paper are presented in the papers themselves (Chapters 4, 5 & 6).

Chapter 5 presents the first paper for this integrated thesis: "Language matters: PIRLS reading achievement in South Africa". This paper provides the foundation for the research by presenting a detailed exploration of reading achievement in South Africa. It aims to understand how different contextual and linguistic factors influence students' PIRLS 2016 reading achievement in South Africa, with a specific focus on differences in achievement across test language groups. In this way, subsequent analyses of the PIRLS 2016 instruments that are undertaken in Papers 2 and 3 could be conducted with an understanding of the role that both contextual and linguistic factors play in the reading achievement of South African students.

1. Paper 1 addresses three research questions:
    1.1. To what extent do contextual factors impact students' overall PIRLS 2016 reading achievement in South Africa?
    1.2. To what extent do factors related to the use of the test language at home impact students' overall PIRLS 2016 reading achievement in South Africa?
    1.3. To what extent does the test language impact students' overall PIRLS 2016 reading achievement in South Africa?

This paper contributes towards the overarching aim of this research by presenting a detailed picture of PIRLS reading achievement in South Africa. It undertakes a multiple linear regression analysis to explain the relationship between test language and reading achievement by considering other potentially influential factors from students' home and school contexts.

The findings from this paper indicate that language in which students wrote PIRLS Literacy 2016 had a substantial influence on their reading achievement scores when controlling for other factors such as their use of the test language at home, socioeconomic indicators such as home resources for learning, students' attitudes towards learning, and factors related to school climate. Notably, students in all nine African language groups had significantly lower scores compared to students in Afrikaans and English groups. The effect of test language is particularly evident for students in the Sepedi language group. These findings motivated further investigation into the psychometric properties of the test itself and the comparability of test and item difficulty across language groups, as well as the selection of the three language groups in Papers 2 and 3.

Chapter 6 presents Paper 2 of the integrated thesis: "Fit for purpose or off the mark: PIRLS 2016 in South Africa". This paper focuses on the psychometric properties of PIRLS Literacy 2016. It evaluates whether PIRLS is able to meet the needs of the South African context by providing valid and comparable information about reading ability across language groups. It focuses on a subset of data from three of the 11 language groups; the two highest performing groups – Afrikaans and English – and the lowest performing group – Sepedi (more details on the selection of language groups is provided in Section 4.3). The analysis in this paper explores the fit of the psychometric models used, checks the assumptions of the models, and evaluates the alignment between student ability and item difficulty across language groups.

2. Paper 2 addresses two research questions:
   2.1. To what extent does the model used by PIRLS 2016 fit students' data from the Afrikaans, English, and Sepedi groups in South Africa?
   2.2. To what extent is the difficulty level of items in PIRLS Literacy 2016 suitably targeted to the ability of students in the Afrikaans, English, and Sepedi groups in South Africa?

Paper 2 contributes to the overarching aim of this thesis by providing a comprehensive analysis of the appropriateness of the PIRLS 2016 instrument for South African students in terms of both the fit of the analytic models used and the targeting of the item difficulty. In this way, subsequent investigation into the comparability of items conducted in Paper 3 is undertaken in a way that accounts for some of the potential threats introduced by the low overall performance of South African students relative to other participating countries.

Chapter 7 presents the third and final paper that makes up this integrated thesis: "DIF happens: The impact of differential item functioning across language versions of PIRLS 2016 in South Africa". It examines the degree of comparability across the three selected language versions, with a specific focus on differences in item difficulty and the overall impact of those differences across groups. This paper applies differential response functioning (DRF) techniques to evaluate comparability at item, passage, and overall test levels.

3. Paper 3 addresses three research questions:
   3.1. To what extent are the Afrikaans, English, and Sepedi versions of PIRLS Literacy 2016 comparable at an item level?
   3.2. To what extent does the comparability of PIRLS Literacy 2016 across these languages at an item level impact on the comparability of the assessment at a passage level?

3.3. To what extent does the comparability of PIRLS Literacy 2016 across these languages at item and passage level impact on the comparability of the assessment at test level?

Paper 3 contributes to the overarching aim of the research by tackling the critical issue of comparability in PIRLS 2016 across language versions. It provides an analysis of potential comparability concerns at the item level as well as quantifying the overall impact of these issues to understand what the differences may mean for overall score comparability across groups.

Chapter 8 builds on key findings from all three papers to contribute towards a discussion of the validity and comparability of PIRLS Literacy 2016 for understanding reading achievement in South Africa. It draws together insights from each paper, as well as relevant literature and other research in the field to unpack the concerns and opportunities of PIRLS Literacy 2016 as an assessment of reading in South Africa. This discussion chapter also provides recommendations for improving assessment programmes to ensure they are suitable for the complex multilingual South African context.

Chapter 9 is the final chapter in this thesis, it reflects on implications of the findings from this thesis in the broader context of large-scale assessment in other low-performing and developing countries. This chapter includes a discussion of the limitations across all three papers, recommendations for future research, as well as the central contributions that the investigations undertaken in this thesis make to both research and practice for the field of ILSA.

CHAPTER 2 CONTEXTUAL CONSIDERATIONS

The influence of contextual factors, curricula, and educational policies on opportunities to achieve academically is widely acknowledged (Crowe et al., 2009; Lagravinese et al., 2020; Mohohlwane, 2019; Ngobeni et al., 2023). The challenge in South Africa is that decades of unequal, segregation-based policies have left in their wake a situation where race, language, home resources, socioeconomic status, and school quality are almost impossible to disentangle. Attempting to understand the results and applicability of a reading test like PIRLS, therefore, necessitates a detailed consideration of both past and present contextual matters as they relate to the educational system in South Africa. This chapter presents an overview of the historical and current educational policies in South Africa as well as a discussion of their ongoing impact on inequalities in the schooling system. It also provides a review of some of the major large-scale assessment studies conducted in South Africa. Finally, it presents a detailed description of South Africa's participation and achievement in PIRLS to date, as well as a discussion of the status and influence of PIRLS in South Africa.

2.1 Apartheid education policies

South Africa has a recent history of systemic oppression on the basis of race. The discriminatory education policies of the apartheid regime systematically and intentionally reduced access, quality and opportunity for black students to learn to read. The atrocities of the unjust policies made on the basis of race during the apartheid regime are highlighted by the words of Hendrik Verwoed (former prime minister) during a speech in 1950:

> There is no place for [the black community] in the European community above the level of certain forms of labour ... What is the use of teaching the [black] child mathematics when it cannot use it in practice? That is quite absurd. Education must train people in accordance with their opportunities in life, according to the sphere in which they live.

During the apartheid regime, education was governed by separate education departments, each of which catered to one of the four hierarchical racial categories constructed by the apartheid government (black, coloured, Indian, and white). The policies and systems followed by these departments were outlined in four different education acts[3]. These policies sought to segregate and disempower on the basis of race, and the curricula and budget allocation differed drastically across racial categories, with government spending at least 10 times more on white students than they did on black students (Bond, 2003). This means that infrastructure, the level of education offered at schools, quality of teacher training, and the budget allocations for students from different racial categories were all prescriptively different.

One of the apartheid educational policies that continues to have a severe impact on educational outcomes relates to the medium of instruction policy (Eriksson, 2014; Mohohlwane, 2019). During apartheid, mother-tongue instruction was mandated for black students for the first five years of primary school, after which the medium of instruction switched to English or Afrikaans (Reagan, 1987). While the merits of an early grounding in mother-tongue language skills for educational success are recognised, this policy was instrumental in ensuring continued segregation and disempowerment on the basis of race. The concerns regarding policies surrounding the medium of instruction were amplified when, in 1974, a new policy mandated that Afrikaans become the medium of instruction for all students in secondary schools across the country.

Resistance to apartheid education policies and practices was building in a variety of platforms including schools and tertiary institutions (Naidoo, 1990). Triggered by the language policy change, these resistance efforts became stronger, and protests began in Johannesburg in June 1976. On the 16th of June, more than 15,000 students came together in Soweto to protest their concerns regarding apartheid education policies and, in particular, the switch to Afrikaans as the compulsory medium of instruction (Ndimande, 2013). This resulted in country-wide confrontations between protesting students and police, during which many protestors (a majority of them black teenagers) were killed by apartheid police. This devastating event highlights the influential and political nature

[3] the Bantu Education Act 1953; the Coloured Persons Education Act 1963; the Indian Education Act 1965; and the National Education Policy Act 1967. (The inherent hierarchy of these separate policies is evident not just in the regulations within the policies but even in the use of 'National' for the act that applied only to white South Africans)

of language policy as a means of oppression that was present during the apartheid regime. The dynamics of balancing language rights, healing past injustice, and providing equitable opportunities for all students is something that South Africa's current education policies are attempting to address.

2.2 Current education policies, systems and practices

Since the fall of the apartheid regime in 1994, South Africa has dismantled all formally racist laws and policies, with new legislation and education policies aimed at redressing past cruelties and promoting social integration across all race groups. Nevertheless, the extreme injustice of apartheid remains pertinent in the current educational sphere in South Africa. Despite efforts to redress past segregation policies, many researchers have shown that student outcomes in South Africa can be largely attributed to three main factors: historical racial categorisation, socioeconomic status, and language of instruction (Fleisch, 2008; Gustafsson & Taylor, 2022; Spaull, 2013; Taylor & Yu, 2009; van der Berg, 2007; van der Berg et al., 2016). The historical discrimination on the basis of race and language in South Africa means that these three factors overlap substantially at the school level (Mohohlwane, 2019). This has resulted in what researchers in South Africa describe as a 'bimodal' structure of schooling, where students that attend formerly advantaged schools continue to achieve considerably higher educational outcomes than those that attend formerly disadvantaged schools (Fleisch, 2008; Spaull, 2013; Taylor & Taylor, 2013; van der Berg, 2007; van der Berg & Gustafsson, 2019).

Added to this, ongoing economic inequalities mean that the current structure of schools in South Africa continues to echo the unequal opportunities that characterised apartheid education. On the one side, there is a system of well-resourced, functional schools that provide quality instruction and facilities. Typically, these schools formerly served either white or Indian (and sometimes coloured) students under the apartheid government, however, they now have more racially diverse student bodies (Taylor & Taylor, 2013). These schools usually charge fees and serve predominantly middle-class students in urban and suburban areas.

On the other side is a system of under-resourced, largely dysfunctional schools that lack basic infrastructure (such as running water, electricity, and toilets) as well as school resources (such as libraries, textbooks, and sufficient teachers). These disadvantaged schools formerly served black or coloured students under the apartheid government. Importantly, while formerly white schools have seen some integration of students from different races, enduring socioeconomic and spatial divisions mean that there has been almost no integration of white students into historically black schools (Taylor & von Fintel, 2016). These schools charge either minimal or no fees and serve mainly poor students in townships, rural areas, and villages.

2.2.1 Language in education policies

Current language-in-education policies intersect with this underlying bimodal system and disparities in student performance are seen when considering the language of instruction at the school. There are currently five key national policies[4] related to language use in education which, together, aim to support the Constitution (1996), heal past divisions, promote social justice, support the right to language choice, and encourage an additive approach to multilingualism.

While language use in schools during apartheid was heavily prescriptive, current policies allow for any of the 12 official languages[5] to be selected as a medium of instruction in the primary school. They recommend a 'home language' to be used in the first three years (while learning English as a subject) before switching to English instruction in Grade 4. School governing bodies (typically comprised of parents, some staff members, and the school principal), and provincial heads of

---

[4] National Education Policy Act No. 27 (1996); South African Schools Act No. 84 of (1996); Language in Education Policy (1997); Curriculum and Assessment Policy Statements (2011); the Incremental Introduction of African Languages in South African Schools (2013)

[5] Although there are currently 12 official languages in South Africa, PIRLS 2016 was conducted in 11 languages as the 12th official language is South African Sign Language, which has only recently received approval to become an official language (Parliament of the Republic of South Africa, 2023)

departments share authority for selecting languages in each school (DBE, 2011a). However, these decision-makers operate within the constraints of the prevailing schooling system. This means that, in practice, clear patterns have emerged in schools in South Africa in terms of language provision along the lines of former racial divides.

Disadvantaged schools in townships, rural areas and villages tend to instruct in one of the nine official African languages in the first three years, which typically reflect the dominant linguistic demographics in their communities. However, in Grade 4 these schools transition to English medium instruction. By contrast, in advantaged schools in suburban and urban areas, English or Afrikaans is often selected (or merely sustained) as the medium of instruction throughout primary school. Despite the racial and linguistic diversification of their student bodies, several complex factors enable the continued legacy of English or Afrikaans instruction in these schools. First, the linguistic heterogeneity of urban and suburban areas in South Africa means that selecting one 'dominant' language can be challenging. Consequently, given the status of English as a language of access and opportunity, decision-makers often prefer English instruction (Taylor & von Fintel, 2016). Additionally, many of these schools have established practices, teachers, materials, principals, and structures that support the continued use of English or Afrikaans throughout the primary school. Using English or Afrikaans from Grade 1 also means that students attending these schools do not undergo a switch in medium of instruction in Grade 4.

Another important issue related to the use of English or Afrikaans as a medium of instruction relates to the language(s) used for initial teacher education programmes at higher institutions in South Africa. Although most universities in South Africa have multilingual language policies, it is rare for degree programmes to be offered in any language other than English (or in some cases Afrikaans) (Makalela, 2016; 2017; Ngubane & Makua, 2022). This situation presents a substantial challenge, as there is a pressing need to train teachers to teach in African languages to meet the needs of the current language in education policies that advocate for mother-tongue instruction in the first three years (Mohohlwane, 2019; Omidire & Ayob, 2022; Spaull & Pretorius, 2019). The disconnect between the languages used for teacher training and classroom instruction creates a barrier to effective teaching and learning in these early years.

Furthermore, concerns have been raised about teachers' proficiency in the African languages used as media of instruction and English. Studies have shown that many teachers struggle with academic proficiency in their mother tongue African languages, as well as lack the necessary English language skills and confidence to effectively teach content subjects in English (Mthiyane, 2016; Nel & Müller, 2010). While some initiatives have been implemented to address these issues, such as the introduction of African language courses in some initial teacher education programmes and the development of multilingual resources, these efforts are not yet sufficient to bridge the gap between teacher preparation and the linguistic realities of South African classrooms (Nakidien, et al., 2021).

There is ongoing critique of the current language policies in South Africa, many of the concerns raised centre around the switch to English in Grade 4 (e.g., Broom, 2004; Evans & Nthulana, 2018; Howie, Venter, & van Staden, 2008; Pretorius, 2014; Shepherd, 2018; Zimmerman, 2014). However, concerns about the implementation, efficacy, and the extent to which policy adequately redresses past injustices also feature (e.g., Coetzee-Van Rooy, 2018; Makalela, 2023; Mohohlwane, 2019; Nkadimeng et al., 2023; Nugraha, 2019; Uys et al., 2007). In essence, while national policies aim to promote racial integration, diversity, and multilingual instruction, the legacy of apartheid means that a substantial divide endures in terms of former education department (as a proxy for race), socioeconomic status, or language of instruction (Spaull, 2013). This divide can be clearly seen when considering student outcomes, and similar disparities are seen in numerous metrics (e.g., dropout rates, university entrance) as well as in the results of national, regional and international assessments across subjects and grades (Spaull, 2013; van der Berg et al., 2011, 2016). The next section gives an overview of large-scale assessments of reading that have been conducted in South Africa since 1994.

2.3 Large-scale reading assessment in South Africa

Since the introduction of new education policies, a number of large-scale assessment initiatives have been undertaken in South Africa to evaluate progress in various subjects and grades. Some of these assessments were designed and implemented locally, while others have been implemented as part of larger international or regional projects. These assessment initiatives primarily aim to diagnose systemic challenges and monitor progress towards narrowing historical achievement gaps caused by the deliberately unequal education system of the apartheid regime (Zuze & Hannan, 2016). As a result of this aim, most of the assessments undertaken focus on the systemic challenges faced in the context, rather than measuring individual student progress. Although the majority of these assessments measured student performance in multiple subject areas, and South Africa has also participated in large-scale assessments of mathematics and science such as the Trends in International Mathematics and Science Study (TIMSS), this overview focuses on student achievement in reading assessments specifically.

The primary large-scale assessments of reading that have been conducted at a cross-national and national level in South Africa include PIRLS, the Monitoring Learning Achievement (MLA) Project, the Southern Africa Consortium for Monitoring Educational Quality (SACMEQ) surveys, the Annual National Assessments (ANAs), and the Systemic Evaluation studies. There are also two ongoing studies that have been conducted at a provincial level, the Early Grade Reading Study (EGRS) and Systemic Tests.

2.3.1 Cross-national assessment studies

The MLA project, which formed part of a larger international Educational for All initiative by UNESCO and UNICEF, was conducted in 1999 in 50 countries and assessed more than 50,000 Grade 4 students in three areas: life skills, literacy and numeracy (Chinapah et al., 2000). The MLA was the first large-scale educational assessment undertaken in post-apartheid South Africa. Of the 12 participating countries from Africa, South Africa achieved the fifth lowest score in the literacy test (Department of Education, 2008). It was administered in 11 official languages and revealed wide disparities in achievement across groups, particularly between the highest and lowest performing provinces (UNESCO, 2014).

One of the largest regional assessment studies that South Africa has participated in is the SACMEQ surveys, established in 1995 by several African ministries of education to track quality trends in literacy, numeracy, and wider education systems (Ross & Genevois, 2006). South Africa took part in SACMEQ II in 2000, SACMEQ III in 2007 and SACMEQ IV in 2013, which were administered in English and Afrikaans. By SACMEQ III, South Africa had expanded its nationally representative sample to over 9,000 Grade 6 pupils across nearly 400 schools (Govender & Hugo, 2020). The competency levels assessed by SACMEQ II showed the majority of South African students were below average readers compared to other participating countries. Around 27% of Grade 6 students could not read and understand a short passage, and students who always spoke English at home scored almost 40 points higher in reading than those who did not (Spaull, 2013). The poor literacy results of SACMEQ are particularly concerning when considering that students were assessed after six years of formal schooling but many could still not demonstrate functional literacy skills. While SACMEQ aims to inform education policy recommendations, some critics have argued that the assessments lack accessible technical documentation and an explicit conceptual framework which has constrained the transparency expected of large-scale, cross-national assessments (Gustafsson & Crouch, 2018; Howie, 2022).

South Africa is one of the 12 confirmed participants planning for the implementation of the fifth cycle of SEACMEQ (which has now expanded to include Eastern African countries) (SACMEQ, 2024). However, the COVID-19 pandemic has meant that data collection shifted from an original 2020 schedule to December 2022 and the overall project timeline extended to 2024.

2.3.2 National assessment studies

There have been two national assessment studies conducted in South Africa that were implemented in 11 official languages; the Systemic Evaluation studies administered in 2001, 2004 and 2007, and the ANAs implemented every year between 2011 and 2014 (Department of

Education, 2003; Kanjee & Moloi, 2014). Both the Systemic Evaluations and the ANAs were designed and administered by the national Department of Basic Education and included evaluations of language and numeracy skills across various grades in both primary and high school.

The results of the first Systemic Evaluations shocked the public with students showing very poor performance in all areas, moreover, progress deteriorated between 2004 and 2007 (DBE, 2010). In 2007, the results for literacy revealed that only 44% of Grade 3 students were able to demonstrate expected grade-level reading skills in the medium of instruction at their school (Govender & Hugo, 2020). Some of the main concerns centred around the discrepancies in achievement across provinces, with average reading scores for students in the Western Cape substantially higher than those in in Limpopo and Mpumalanga.

The ANAs were introduced in 2011 as a means of monitoring the level and quality of education across all schools after a number of repeated investigations showed that students in South Africa were underperforming in relation to government investment in education (Chisholm & Wildeman, 2013; Kanjee & Moloi, 2014). ANAs were conducted nationally and assessed over 7 million students in Grade 1-6 and Grade 9. The results from ANAs highlighted persistent gaps in reading proficiency across languages, provinces, socioeconomic status, and gender. Unlike the results of other assessments in the country, ANA results showed large increases in average achievement over time for some grades (e.g., a 49% increase in the average for reading in Grade 3 between 2011 and 2012). However, the comparability of these assessments over time and across grade levels has been strongly criticised (Spaull, 2015; van der Berg, 2015).

Indeed, ANAs were a major source of controversy in South Africa on multiple fronts (Govender & Hugo, 2020). The content and design were critiqued for aspects such as the content covered, a lack of variety in the questions, and the application of statistical and procedural good practice was questioned (Spaull, 2015; van der Berg, 2015). Additional criticisms focused on the administrative burden for teachers and students, including issues such as encouraging teaching to the test, and promoting rote learning over deeper understanding (Govender & Hugo, 2020; South

African Democratic Teachers Union (SADTU), 2014). The administration and implementation of the ANAs also faced critique regarding unsatisfactory organisation, a lack of independence in administration and test design, questionable marking practices, and concerns about the reporting of results (Bansilal, 2012; Chilenga-Butao et al., 2020; van der Berg, 2015). Although the ANAs were introduced as part of a long-term national development plan, they were cancelled in 2015 due to pressure from teacher unions and a national assessment programme has not been reinstated (Gustafsson, 2019b; SADTU, 2014).

2.3.3 Provincial assessment studies

More recently, the Department of Basic Education in collaboration with provincial governments, international funders, and academics in South Africa, have introduced some evaluation studies which are being undertaken at a provincial level. The EGRS has been implemented in two provinces of South Africa – North West (EGRS I) and Mpumalanga (EGRS II); and a programme of Systemic Tests has been implemented in the Western Cape.

The EGRS is a series of large-scale evaluations that aim to gather evidence about the effectiveness of intervention strategies to improve early grade reading in South Africa. It assesses reading progress in two languages Setswana (as a home language) and English (as an additional language) (DBE, 2022). Early grade reading assessment (EGRA) tests, which are part of an international initiative originally developed by RTI International and the World Bank and are now implemented in many countries globally, are used to monitor reading progress (Gove & Wetterberg, 2011). Although not implemented nationally, the EGRS has exposed distressing reading results among early grade students in selected provinces, with 70% of Grade 1s unable to sound letters fluently and 83% unable to read a single word in the oral fluency assessment (DBE, 2021). These results vary across home language and additional language subjects; in Grade 2, 58% and 74% of students could not read a single word in Setswana and English assessments respectively. Worryingly, within classrooms there appears a bimodal distribution of reading ability – relatively few students

(13%) could read with partial fluency while the majority still could not decode or comprehend a single word, despite two years of schooling.

Systemic Tests are a provincial assessment programme that has been conducted annually in the Western Cape since 2001. In 2011, these tests were updated to align with curriculum so as to evaluate the success of systemic changes in curriculum as well as teaching and learning strategies (van der Berg et al., 2022). These tests are taken by Grade 3, 6, and 9 students and results reveal similar patterns of low performance in reading as seen in other assessment results. Results from the Systemic Tests did show positive trends in reading progress from 2011-2018, before the COVID-19 pandemic disrupted assessments in 2020 and learning in 2020-2021. A recent report on the impact of the COVID-19 pandemic on learning in South Africa has used evidence from the Systemic Tests, revealing significant learning losses (van der Berg et al., 2022). Students regressed behind previous cohorts by up to 70% of a school year in language subjects. The findings from this report are being used to expand literacy programs and incorporate extra reading, writing and mathematics time to try to claw back pandemic losses.

Overall, findings from large-scale assessment studies in South Africa emphasise the need to better utilise assessment data to systematically address inequality in reading outcomes. There are a number of researchers who make extensive use of the data from these studies to understand what matters for reading outcomes in the current educational landscape in South Africa, as well as the effectiveness of intervention programmes (see Spaull & Jansen, 2019; Taylor & Taylor, 2013; Taylor, 2019; van der Berg et al., 2016, 2022; van Staden & Combrinck, 2023). However, although ongoing studies like the EGRS and the Systemic Testing programme, as well as regional studies like SACMEQ, evaluate reading in multiple languages, they do not assess achievement in all official languages of South Africa. This speaks to the important role that PIRLS has in South Africa as it is currently the only measure of reading achievement that assesses reading across 11 languages, as well as measuring progress trends in reading over time.

2.4 South Africa's participation in PIRLS

South Africa has participated in PIRLS since 2006, testing a nationally representative sample of Grade 4 students across all 11 language groups in every cycle. They have also tested Grade 5 (2006-2016) or Grade 6 (2021) as a benchmark sample in each cycle so as to gauge progress between grades, but these benchmarking grades have not always included all language groups. The 2006 cycle was the only one where the Grade 5 cohort included all 11 language groups. In the other cycles, the benchmark samples were limited to Afrikaans and English (2011 & 2021), and Afrikaans, English, and isiZulu (2016).

The remainder of this section, therefore, focuses on the Grade 4 cohorts as they include a nationally representative sample of the population of students across all 11 language groups. PIRLS recommends minimum samples of 4,000 students and 150 classes (Almaskut et al., 2023). However, South Africa oversamples students in order to explicitly stratify across their 11 official language groups. Similar oversampling occurs in other countries (e.g., Canada and Spain), typically oversampling is done by region within countries to ensure representative samples that can be used for subgroup comparisons (see Brochu et al., 2017).

2.4.1 PIRLS reading achievement in South Africa

In every cycle of participation to date, South Africa has achieved the lowest average reading score of all participating countries in PIRLS (DBE, 2023; Howie, Combrinck, Roux, Tshele, et al., 2017; Howie et al., 2012; Howie, Venter, van Staden, et al., 2008). Table 2-1 gives an overview of South Africa's average achievement in PIRLS as presented in the national reports for each cycle, however, before considering these results several important issues need to be specified.

First, in 2006, the achievement of students from African language groups was deemed too low for reliable estimation and as a result only the Afrikaans and English groups were sufficiently robust for trend comparisons (Howie, Combrinck, Roux, Tshele, et al., 2017). This also meant that only the results from the Grade 5 cohort were included in the 2006 international report (Mullis et al., 2007).

Second, due to overall low achievement in 2006, South Africa participated in an easier version of PIRLS in 2011 called prePIRLS. This version of the assessment in 2011 was not equated with the main version of PIRLS at the time. Thus, the 2011 results for prePIRLS participants (South Africa, Botswana, and Columbia) were not comparable with the other 46 participating countries in that cycle. This also meant that benchmarking insights for South Africa across Grade 4 and Grade 5 cohorts were not possible in 2011. This non-equated design explains the incongruence between the 2011 results and the results from other cycles seen in Table 2-1. Additionally, reservations about the reliability of results for South African Grade 5 students in 2011 were flagged in the international report as 15-25% of students had achievement levels that were too low for estimation (Mullis et al., 2012).

Third, although the results from prePIRLS 2011 were retrospectively rescaled in 2016 to enable cross-cycle trend comparisons, South African researchers disputed the rescaling methodology that was used (Gustafsson, 2020; Gustafsson & Taylor, 2023). These disputes were recognised by the IEA in March 2020 who noted on their website, "Due to problems with the conversion applicable to 2011, IEA no longer considers the published 2011 to 2016 trend specific to South Africa reliable" (IEA, 2020). This resulted in the removal of South Africa's trend comparisons for 2011-2016 from the PIRLS international report, and a revised trend analysis was conducted by the IEA (2021). However, although an unpublished document detailing this analysis was shared with the Department of Basic Education in South Africa, the updated trend has not been republished publicly and trend results between 2011 and 2016 remain absent from the PIRLS website (Gustafsson & Taylor, 2022). In 2016, South Africa also participated in an easier version of the main test, which was renamed PIRLS Literacy and is scalable with the main test. Nevertheless, the problems with the scaling between 2011 and 2016 mean that one of the central benefits of consistent participation in PIRLS – monitoring long-term progress across cycles – is no longer possible for South Africa.

Finally, similar to concerns raised in 2006 for students who took the test in African languages, the IEA noted reservations about the reliability of PIRLS 2021 results for South Africa because more than a quarter of students had achievement “too low for estimation” (Mullis et al., 2023, p. 28). In 2021, these concerns pertain to the full sample and not just the students who took the test in an African language as in 2006. The international report does not elaborate on the exact proportion of students who had achievement too low for estimation in South Africa, what the cut off criteria were, what these reservations mean for interpreting the results, or how these concerns affect inferences that can be drawn from the PIRLS 2021 data. Notably, PIRLS 2021 did not include a separate ‘easier’ option. Instead, a group adaptive design was used where countries chose ratios of easy, medium, and difficult passages based on their results from previous cycles. South Africa had a ratio of 30% difficult and 70% easy passages, as recommended by the IEA for countries with average performance below 450 points (DBE, 2023). This approach was adopted to enable more reliable comparisons across the wide ranges of achievement seen between countries, particularly to expand coverage at the higher end of the ability distribution which was not addressed through previous designs (Martin et al., 2019).

The average reading scores and low benchmark achievement for South Africa in each cycle of PIRLS are shown in Table 2-1. These results are drawn from the national reports published after each cycle (i.e., DBE, 2023; Howie, Combrinck, Roux, Tshele, et al., 2017; Howie et al., 2012; Howie, Venter, van Staden, et al., 2008). However, due to the non-equated design of prePIRLS 2011, the initial 2011 average of 461 points presented in the first national report was not comparable to other countries or other cycles for South Africa. The rescaled 2011 results presented in the international report for PIRLS 2016, as well as the revised scores presented by Gustafsson (2020) and IEA (2021) are shown in the table notes.

Table 2-1

Overview of Grade 4 PIRLS results in South Africa 2006-2021

Cycle	Number of students	Average reading score	% did not reach lowest benchmark	Gap between highest and lowest performing language groups
2006	16,073	253	87%	-*
2011	15,744	461[a]	29%[a]	137[a] points (Afrikaans vs Sepedi)
2016	12,810	320	78%	96 points (English vs Sepedi)
2021	12,426	288	81%	176 points (Afrikaans vs Setswana)

Sources: National reports for PIRLS 2006 (Howie et al., 2008); 2011 (Howie et al., 2012); 2016 (Howie et al., 2017); and 2021 (DBE, 2023)
Notes:
** Cross-language comparisons not reliable in 2006 due to low achievement in African language groups*
[a] The initial 2011 score presented in the PIRLS 2011 International report published in 2012
Rescaled scores:
b. 323 presented by the IEA in the initial PIRLS 2016 international report (Mullis et al., 2017a)
c. 295 presented by Gustafsson (2020) in a reanalysis of the PIRLS 2011 data disputing (b)
d. 278 presented by the IEA (2021) in an unpublished reanalysis of the 2011-2016 trend

South African students have consistently achieved average reading scores well below the international centerpoint (500) since first taking part in 2006. Moreover, only 13-22% of South African Grade 4 students have reached the low international reading benchmark in PIRLS assessments. This means a majority of students could not complete basic reading tasks like retrieving explicitly stated details from a text. Added to this, the distribution of achievement in South Africa has been the widest of all participating countries, with substantial gaps persisting between language groups in South Africa. For example, in 2021 the average score for Afrikaans (387 points) was 176 points higher than the average for Setswana (211 points). Such enduring disparities and overall low achievement underline South Africa's literacy crisis.

While South Africa's overall PIRLS achievement remains extremely worrying, and concerns about the reliability of comparisons over time have been raised, educational stakeholders continue to find value in participation by investigating micro-level growth signs and trends wherever possible. The 2021 PIRLS national report states that, despite reservations around macro-level data comparisons, they hope that deeper investigations into the data from PIRLS 2021 "will provide meaning to and justification for" South Africa's participation in PIRLS (DBE, 2023, p. 4). It is therefore essential to ensure that concerns about PIRLS data are thoroughly investigated, particularly because results and information from previous PIRLS cycles continue to have an influence on policy decisions and interventions within the country.

2.4.2 Status of PIRLS in South Africa

PIRLS has been noted as one of the leading instruments for monitoring progress towards UNESCO's Sustainable Development Goal (SDG) 4 – Quality education for all – internationally (UNESCO & IEA, 2017). South African national reports on PIRLS add further detail on the instrumental role PIRLS can play in realising the specific targets of SDG 4 within the country (DBE, 2023). The contextual questionnaire data from the PIRLS student, home, school, and teacher questionnaires is noted as beneficial for assessing the availability of early childhood facilities (Target 4.2 – Early childhood), gauging students' attitudes about their learning environment (Target 4a – Learning environment), obtaining information from teachers about pedagogy and school climate (Target 4c – Teachers). Furthermore the Department of Basic Education (2023) highlights that, by testing in 11 official languages, PIRLS can supply crucial evidence about comprehension skills across languages to guide reforms of language in education policies, inform targeted resource provision, and support equitable language development initiatives targeting the groups and areas with the greatest literacy needs (Target 4.5 – Equity).

Since first participating in 2006, South Africa's PIRLS results have frequently been the focus of public, governmental, non-governmental, and academic discourse about reading achievement and required reforms. Consequently, there have been multiple publications which have focused specifically on PIRLS in South Africa in the form of research reports (e.g., Gustafsson, 2020; Taylor & Taylor, 2013; van der Berg et al., 2016), journal articles (e.g., Combrinck & Mtsatse, 2019; Janks, 2011; van Staden & Howie, 2012, 2014; Zimmerman & Smit, 2014), media reports (e.g., Andersen, 2017; Maimane, 2023; SAHRC, 2023; Taylor, 2017), as well as book chapters (e.g., Spaull & Pretorius, 2019; van Staden & Gustafsson, 2022). While some of these publications critique the assessment's local relevance and value, many investigations highlight constructive ways results have informed policies targeting inequality.

For instance, examining early impacts, Howie and Venter (2012) detail the instrumental role 2006 findings played in exposing reading deficits and catalysing governmental prioritisation of initiatives to improve reading development. Some of the initiatives implemented shortly after PIRLS 2006 results mentioned in their chapter include new national reading campaigns, increased distribution of storybooks in all 11 languages, the release of a national reading strategy document, and the doubling of the public library budget (Howie, 2012; Howie et al., 2012). In a comprehensive overview of reading policy reform, Spaull and Pretorius (2022) show how the alarmingly low achievement exposed in early PIRLS cycles helped confirm reading as a top priority, while results from later cycles showing ongoing gaps across languages further focused efforts to improve reading for all students. Finally, a recently published book edited by van Staden and Combrinck (2023), *Tracking Changes in South African Reading Literacy Achievement,* reflects on lessons learned from a decade of participation in PIRLS and presents further analyses of the data to add additional insights into the situation. These are just some examples of studies that have shown that, despite persisting doubts about the reliability of the trends and applicability of the assessment for the local context, the data continues to inform policies in the country.

Although the direct influence of ILSAs results on policy can be difficult to gauge (Rutkowski et al., 2020); the impact that PIRLS has had on the monitoring and evaluation of national development targets is reflected in changes made to South Africa's newest national development goals. In 2019, President Cyril Ramaphosa announced in his state of the nation address that one of the top five development goals for the next decade in South Africa is that "every 10-year-old child will be able to read for meaning" (Ramaphosa, 2019). In the previous National Development Plan (released in 2010) reading achievement did not feature in the top 10 priorities directly (National Planning Commission, 2010). The only milestone included in the plan that was explicitly related to reading achievement was: “Increase the quality of education so that all children have at least two years of preschool education and all children in Grade 3 can read and write” (p. 34).

The specific impact of PIRLS on South African development goals is evident when comparing the word choice in both statements. Two major changes to the milestone from 2010 to the goal in 2019 may be directly related to PIRLS results in South Africa. Firstly, the change from in ‘Grade 3’ to ‘10-year-old’; in South Africa a Grade 3 child would typically be 8 or 9 years old and in their third year of formal schooling. PIRLS requires participants to be in their fourth year of formal schooling. Therefore, by changing to age 10, it would be possible to use PIRLS as a method of monitoring and evaluating the progress towards this national development goal. The second notable change is the shift from “can read and write” in 2010 to “read for meaning” in 2019 – a word choice that reflects the way in which the PIRLS reading framework highlights the construction of meaning as a central component of reading.

One of the most recent examples of the influence of PIRLS on national policy in South Africa is the adoption of the PIRLS reading framework as the foundation for a new national framework for assessing reading comprehension. In 2021, the Minister of Basic Education, Angie Motshekga, announced that the PIRLS reading framework would serve as the foundation for a unified national framework aimed at enhancing the understanding of reading comprehension (Motshekga, 2021). The theoretical and pragmatic motivations for using the PIRLS framework are outlined in a report by the South African Human Rights Commission (SAHRC, 2021). According to this report, the four comprehension processes outlined by the PIRLS framework make it manageable yet nuanced enough to identify higher-order reading skills. They also add that transparency of the PIRLS methodology and procedures, together with South Africa's previous rounds of participation, provides great potential for gaining insights into reading development across all 11 languages. The decision to take on the PIRLS framework reflects the extent of the influence that PIRLS continues to have on reading reform in South Africa. This further emphasised by the use of PIRLS data by both the

Department of Basic Education and the South African Presidency as a formal means of monitoring progress towards national goals (SAHRC, 2021).

The influence of ILSAs on policy in other African countries has also been shown. Mulongo and Amod (2019) analysed strategic documents in South Africa, Kenya, and Tanzania, finding more than 10 reforms between 2000-2015 that responded directly to cross-national assessment recommendations, including from PIRLS. Other investigations into the influence of PIRLS on national policy internationally have brought to light concerns regarding the relevance of PIRLS data to local contexts, specifically when policy decisions were being made (Lenkeit & Schwippert, 2018; Popat et al., 2017; Gomendio & Wert, 2023; Volante et al., 2022). These concerns point to the need for additional detailed, critical analyses of PIRLS that interrogate its relevance and applicability in the diverse contexts in which it is used. This is particularly true in South Africa where the influence and status that PIRLS has had on education, coupled with the unique and complex inequality in schooling in the country, has implications for interpreting and analysing the results.

2.4.3 Situating PIRLS results in the context of South African schools

In assessing and interpreting student performance considering South Africa's current education system is essential. The achievement of students cannot be understood without considering how factors such as race, socioeconomic status, language of instruction, home language, and school quality intersect within the highly unequal system of schooling. Students attending advantaged schools drastically outperform those in disadvantaged schools (Fleisch, 2008; Spaull, 2013; van der Berg & Gustafsson, 2019). Moreover, the majority of students in South Africa (between 75-85%) fall into the disadvantaged school category (Jansen, 2019). Consequently, national averages on large-scale assessments like PIRLS mask the severe inequalities across the two subsystems of schooling as they are skewed by the top 15-25% of students who attend more functional schools (Spaull & Pretorius, 2019).

Complicating matters further, a student's school serves as a proxy for multiple intersecting historical, socioeconomic, and policy factors influencing academic achievement. At the same time, the school a student attends determines their access to critical resources, facilities, textbooks, qualified teachers, and overall learning opportunities. Similarly, language achievement patterns align with historically rooted differences in schools' language policies and medium of instruction provisions. This means that comparing achievement across student groups risks conflating issues of language, historical disadvantage, and current school quality. It is therefore essential to situate any subsequent analyses of the data from these tests (such as the investigation being undertaking in this thesis), within the context of a separate and unequal schooling system that continues to reflect apartheid-era divisions in South Africa.

CHAPTER 3 LITERATURE REVIEW

As an integrated thesis, each paper contains a focused review of the literature tailored to its specific research aims and analysis. However, before delving into the three papers, it is essential to establish a comprehensive understanding of the fundamental concepts and challenges surrounding the assessment of reading comprehension in diverse linguistic contexts, as they inform the work presented in this thesis. This chapter explores the theoretical frameworks related to the assessment of reading comprehension, which is the primary construct measured by the PIRLS instrument, and reviews literature related to challenges with assessing reading comprehension across languages. This chapter also discusses the key components of quality assessment, namely validity, reliability, and fairness as these have implications for the strength and meaningfulness of conclusions drawn. Furthermore, this chapter provides an overview of some of the critical concerns that threaten the quality of PIRLS in the context of South Africa. By presenting this foundation, this chapter aims to establish a shared understanding of the fundamental concepts and considerations that permeate the three papers.

3.1 Considering the construct

The effective assessment of any area of student academic performance relies on quality instruments that can accurately measure the intended construct. Since this thesis scrutinises PIRLS, it is important to establish a shared understanding of the construct it assesses – reading comprehension. Reading comprehension is a multifaceted cognitive process that involves decoding written text and constructing meaning from it. Approaches to assessing this construct have evolved considerably over the years, driven by advancements in the theoretical and conceptual understanding of reading as well as the refinement of measurement techniques. This section provides an overview of some of the major theoretical frameworks that have shaped our understanding of reading comprehension, as well as the influence they have had on how reading is assessed.

3.1.1 Theoretical frameworks of reading comprehension

Early frameworks for understanding reading comprehension included both bottom-up models and top-down models. Bottom-up models saw reading as a process of gradually gaining understanding by decoding smaller units into larger constructs (Gough, 1972; LaBerge & Samuels, 1974). Readers were thought to process written symbols, recognise word patterns, parse syntactic structures, and integrate this information in a sequential way to incrementally comprehend a passage. As a result of these views, early reading assessments were focused heavily on oral reading fluency and accuracy of word decoding (Ebel, 1972; Paris & Stahl, 2005).

Meanwhile, top-down models suggested that higher-level knowledge was central to facilitating comprehension (Berger & Perfetti, 1977; Perfetti & Goldman, 1976). In these models, reading comprehension was seen to rely substantially on readers' prior experiences, language skills, and abilities to use existing background knowledge during the construction of meaning efficiently. Interactive models integrated these processes, demonstrating that meaning making during reading emerges from an interplay between deciphering text and applying background knowledge (Rumelhart, 1977; Stanovich, 1980).

As theoretical perspectives began to recognise the complex, interactive, and constructive nature of reading, assessments shifted their approach and aimed to capture information about comprehension skills rather than just word-level skills (Cain & Oakhill, 2006; Sarroub & Pearson, 1998). Standardised reading tests became more widely adopted from the 1970s, and although they were broadly criticised for issues with validity and cultural bias, they formed a foundation for a rapid evolution in assessment techniques (Popham, 1999). A majority of these early versions of standardised reading tests sought to understand and unpack students' cognitive and reading skills (Pearson & Hamm, 2005).

The Simple View of Reading (SVR) then emerged which further emphasised that reading comprehension is dependent on decoding skill and language comprehension abilities (Gough & Tunmer, 1986; Hoover & Gough, 1990). The SVR suggests that decoding skills allow printed words to be recognised while language comprehension provides the vocabulary knowledge needed to interpret the meanings of words being read. This hypothesis maintains that both word-level reading skills and higher-order comprehension of language are essential to overall literacy development and proficiency. The SVR has given rise to a number of assessments that focus specifically on understanding the components involved in reading by looking at the extent of students' vocabulary knowledge, and their decoding skills. Some examples include the Peabody Picture Vocabulary Test which measures receptive vocabulary, or the ability to understand spoken or written words (Dunn & Dunn, 2007); and the Woodcock-Johnson Achievement Test which has subtests focused specifically on vocabulary and decoding abilities (Woodcock et al., 2001). Research drawing on the SVR framework often makes use of tests such as these to examine the distinct contributions of vocabulary knowledge versus decoding skill to overall reading comprehension (e.g., Goriot et al., 2021; Kang & Shin, 2019; Kim, 2017; Vaughn et al., 2020).

At the same time, sociocultural frameworks began to provide further insight into the processes and context involved in making sense of text, these frameworks focused not just on the basis of the skills required to read but also incorporate the ways in which the understanding of language and texts is shaped by cultural context (Gee, 1989; Heath, 1982; Johnson & Perry, 2012). Sociocultural frameworks highlight that reading is an active process of constructing meaning; highly responsive to shared literacy practices, power dynamics, and intended purposes situated within particular communities (Perez et al., 2004). As interest grew regarding the broader systemic influences at play in academic achievement, cross-national achievement assessments and surveys began to emerge (Kamens & McNeely, 2010). These larger-scale assessments went beyond just assessing student outcomes and sought to understand and evaluate the effectiveness of educational systems across different contexts more holistically (Volante et al., 2022).

The influence of context on constructing meaning in reading is widely discussed, with theorists noting that contextual influence can be both beneficial and detrimental to reading comprehension (Christianson & Luke, 2011; Cummins, 2012; Miller & Faircloth, 2009; Rueda, 2013).

Developing insight into the interaction and impact of the various skills and factors required for successful reading development has been a key feature of reading research (Castles et al., 2018; Goldenberg, 2020; Ouellette & Beers, 2010; Snowling et al., 2019). Empirical investigations have revealed that proficient reading depends on the coordination of multiple factors which can be broadly categorised into individual student characteristics, features of the home environment, and factors related to teaching and learning at school (Hulme et al., 2022; Melby-Lervåg & Lervåg, 2014; Westerveld et al., 2020). Understanding the interplay of these various factors is crucial for the present study, which aims to understand interactions between the South African context and the assessment of reading.

3.1.2 Challenges with assessing reading comprehension across languages

One of the key features of ILSAs is that they have versions in different languages, PIRLS 2016 for example was conducted in 50 countries and translated into over 40 different languages (Mullis et al., 2017a). While carefully constructed assessments can provide valid insights into reading ability, challenges remain, particularly when attempting to assess reading across multiple language versions. Isolating reading skills from language and conceptual knowledge is an extremely complex undertaking, particularly because language is both part of the construct being measured and the medium through which the assessment takes place. This challenge is further compounded by the complex and dynamic multilingual contexts that exist in many countries.

Much of the early research into reading development focused on learning to read in English, either as a first or, less frequently, as a second language. Where multilingualism was considered in the development of reading, it was mostly investigated in terms of the way in which fundamental reading skills were transferred from the students' first language to facilitate reading development in a second language (Alderson, 1985; Bernhardt, 1993; Goodrich et al., 2013; Vermeer, 2004). Although these studies provided key insights into the influence that language-specific features such as orthography, morphology, and phonology can have on the development of reading across languages (e.g., Kuo & Anderson, 2006; Pae & Sevcik, 2011; Rajaram, 2018); they do not fully capture the diversity of global language environments where children speak and encounter multiple languages from early ages (Miller et al., 2018).

Contemporary research argues that reading processes must be situated within the complex and evolving multilingual contexts which increasingly characterise schools and societies today (Coady et al., 2022; Kretzer & Kaschula, 2021; Makalela, 2017; McBride, 2017; Murphy, 2018). These perspectives recognise the dynamic interrelations between the linguistic features of multiple languages in a reader's repertoire by considering skills that go beyond linguistic transfer such as multiculturalism and translanguaging (Goldenberg, 2020; Makalela, 2015; Murphy & Unthiah, 2015; Sefotho & Makalela, 2017; van Rooy & Pretorius, 2013; Zulu et al., 2008).

The challenges of learning and assessing reading in multilingual contexts are particularly salient in South Africa. With 12 official languages, many South African students encounter multiple languages from an early age, both at home and in school environments. The linguistic diversity present in South African classrooms means that the reading processes of many students cannot be fully understood through a monolingual lens or by solely considering linguistic transfer from a first language to a second language (Coetzee-Van Rooy, 2018; Makalela, 2015). Added to this, the Language in Education Policy in South Africa requires students to learn to read in a minimum of three languages from their first year of primary school.

Furthermore, sociolinguistic dynamics including language status and cultural-historical relationships between languages can profoundly shape students' opportunities to learn, their attitudes, engagement, and motivation for reading (Boakye, 2015; Casale & Posel, 2011; Makalela, 2023; Mphasha, 2016; Netten et al., 2016; Wang & Wu, 2023). This is particularly true in developing contexts where the colonial languages used at school are often different to local languages, as well as higher in both status and opportunity (Boakye, 2015; McKinney, 2017; Nag et al., 2023; Trudell & Schroeder, 2007). In South Africa, these power differentials between languages can shape students' attitudes, engagement, and motivation for reading in different languages, potentially affecting their performance on reading assessments (Boakye, 2015; Makalela, 2017, 2023). Additionally, the linguistic factors involved in reading assessments may vary across South African languages due to differences in features such as orthography, morphology, and phonology (Malda et al., 2014; Probert & De Vos, 2016; Wilsenach, 2013, 2015; Zulu et al., 2008). These language-specific factors, coupled with the complexities of multilingual language use, mean that there are a number of linguistic factors involved in reading assessment that may affect the difficulty of a test in a specific language which are not necessarily a result of poor translation and may not be easy to identify (El Masri et al., 2016; Nkadimeng et al., 2023).

Developing assessments that effectively measure reading comprehension holistically, therefore, requires test developers to carefully define the specific skills, knowledge and thought processes that constitute reading while also accounting for the influence of the broader social context (Oliveri et al., 2015). This informs the design of comprehensive assessment frameworks that describe the relevant components, constructs, processes, and difficulty levels targeted in a reading assessment (e.g., Mullis et al., 2015a; OECD, 2023).

3.1.3 The PIRLS reading framework

In each cycle, the IEA seeks to incorporate the latest approaches to ensure the framework reflects an up-to-date notion of reading (Mullis et al., 2015b). PIRLS 2016 defines reading as:

> the ability to understand and use those written language forms required by society and/or valued by the individual. Readers can construct meaning from texts in a variety of forms. They read to learn, to participate in communities of readers in school and everyday life, and for enjoyment. (Mullis et al., 2015b, p. 12)

This definition sees reading as a purpose-driven process that is both constructive and interactive. This framework aligns with a number of theoretical frameworks of reading comprehension that suggest skilled reading integrates a number of interrelated factors in the construction of meaning (Dooley, 2017; Huey, 1908; Kintsch & van Dijk, 1978; Laufer, 1989; Thorndike, 1917). The focus on purpose as central to the construct of reading led PIRLS 2016 to concentrate on two overarching reading purposes: reading for literacy experience, and reading to acquire and use information (Mullis et al., 2015b). In addition to reading purpose, the construct of reading assessed in PIRLS Literacy 2016 pays particular attention to four comprehension processes involved in reading: focusing on and retrieving explicitly stated information; making straightforward inferences; interpreting and integrating ideas and information; and evaluating and critiquing content and textual elements.

The PIRLS reading framework also builds on ideas of language comprehension and context being central to reading by incorporating views from Snow (2002, 2018) and Kintsch (2013) that reflect the constructive nature of creating meaning during reading while also incorporating a contextual questionnaire framework that collects background information related to the broader social influences on reading achievement (Hooper et al., 2015). A more detailed discussion of the PIRLS contextual questionnaire framework, as well as the role that background factors play in reading development is included in Section 5.2.

Assessing reading comprehension within this framework, therefore, requires a nuanced approach that acknowledges the dynamic interplay between the various languages in a student's repertoire, as well as the sociolinguistic dynamics at play in a particular context. This is a challenging undertaking and a number of components must be carefully considered and evaluated to ensure the quality of large-scale assessments like PIRLS.

3.2 Considering the components of quality assessment

Effectively assessing any area of student academic performance requires quality instruments that can accurately measure the intended construct and produce reliable, interpretable results that serve the assessment's intended purpose. The purpose of the assessment is a crucial consideration as the extent to which precise and reliable measurements are needed depends on how the assessment results will be used (Newton, 2017). The key foundational components of quality assessments include validity, reliability, and fairness; each encapsulating a wide range of interdependent features that need to be upheld to ensure assessments are producing quality information (AERA, APA & NCME, 2014).

While validity is widely recognised as the most fundamental consideration in assessment, Newton and Shaw (2014) suggest that reliability and fairness are necessary conditions for ensuring assessment quality. According to them, these three concepts are intricately interconnected and when either reliability or fairness is compromised, validity is also undermined. Guided by international standards and practices, ILSAs employ rigorous procedures throughout development and administration to uphold principles of validity, reliability and fairness. However, the expanding diversity of participating countries amplifies threats that require continuous scrutiny across all three components. The sections that follow give an overview of some of the key developments and issues raised in the literature regarding each of these critical components, as well as some concerns that may threaten them in the context of PIRLS in South Africa.

3.2.1 Conceptualising validity

There is extensive literature within the field of educational assessment which has led to sometimes disparate definitions and conceptualisations of validity (Newton & Shaw, 2014). Early researchers established validity as the degree to which the inferences made from test scores can be supported by evidence (APA et al., 1954; Lindquist, 1951). Chronbach and Meehl (1955) proposed different types of evidence that should be used to evaluate validity namely: content-, criterion-, and construct-based evidence. The notion of multiple types of validity evidence gained traction, and progress in validity theory reflected an increasing recognition of its complexity and multifaceted nature (Dunnette, 1963; Guion, 1977; Wernimont & Campbell, 1968). However, discussions continued to seek clarity about the interrelatedness and potential overlap that exist between the different ‘types’ of validity (Guion, 1980; Tenopyr, 1977).

Key shifts were then made through the work of Messick (1989) who proposed a more unified framework of validity which positioned construct validity as the overarching consideration in evaluating assessments. Kane (1992) added to this conceptualisation of validity by providing a framework that links together the series of inferences made when interpreting assessment information, from the initial interpretations of student responses to the decisions and use of inferences made from the final test score. Messick (1995, 1998) went on to highlight the importance of the interpretation and use of test scores when evaluating validity. In other words, validity was not just the concern of a test, but also of the consequences of the interpretations made on the basis of test scores (Messick, 1998). The Standards for Educational and Psychological Testing (*Standards*) recognised this shift and emphasised that test scores need to be validated for each intended interpretation (AERA et al., 1999).

While some dispute the centrality of interpretation to measurement – such as Borsboom et al. (2004) and Borsboom (2005), who argued that the interpretation of test scores is secondary to establishing a sound measurement model that accurately represents the attribute being measured – most agree that the way scores are interpreted and used is crucial for establishing validity. The prevailing perspective maintains that validity hinges on the interpretations and use of test scores for their intended purposes. In this perspective of validity, researchers stress that validating a test needs to be done for each intended purpose separately (Cook & Beckman, 2006; Downing, 2003; Kane, 2001).

Newton (2007) highlighted the importance of clarifying the different purposes of educational assessments, both primary and potential secondary purposes, so that tests can be designed in a way that enables sufficiently valid inferences to be drawn from the results. Central to this is the idea that the context in which a test is used and the characteristics of the groups taking the test do not just impact the validity of the results, but also shape the overall validity argument and the kinds of evidence required to support it (Embretson, 2007; Hambleton et al., 2004; Sireci, 2007). This view of validity as something that is both empirically evaluated (in terms of the test's construct) and value-laden (in terms of the test's context, purpose and use) has become central to subsequent interpretations of validity that continue to emerge (Haertel, 2013; Kane, 2013; Newton, 2012; Newton & Shaw, 2014). More recent developments synthesise the earlier notions of validity to present an integrated approach that incorporates test content, student responses, analytic methods, contextual variables, as well as the social and political consequences of test interpretation (Hughes, 2018; Knoch & Chapelle, 2017; Reynolds et al., 2021).

This multifaceted concept of validity is particularly challenging to uphold in ILSAs due to the extensive array of contexts in which they are used, and the increasing range of uses to which they are put (Addey et al., 2020). Establishing validity for the primary purpose of measuring and comparing educational achievement across countries is already a massive undertaking that requires rigorous evaluation of construct representation, ensuring appropriate measurement models, and reducing potential for any misuse of the results. However, the increasing range of secondary uses of ILSA data for within-country policy decisions, resource allocation, and interventions adds further complexity to this challenge. Although ILSAs undergo evaluation to ensure validity for their primary purpose – to provide comparable results across participating countries, the variety of secondary purposes for which ILSAs are used often do not receive equal scrutiny during the test development process (Addey et al., 2020; Newton, 2017; Reynolds et al., 2021). As large-scale assessments expand into new settings and uses, therefore, thoroughly scrutinising the continuing validity of interpretations and applications of tests in new, diversifying contexts is essential (Addey et al., 2020; Oliveri et al., 2018). This is a particularly important consideration when evaluating validity as providing comparable information so as to inform educational policy and 'best practice' form part of the core aims of ILSAs like PIRLS (Martin et al., 2019; Mullis et al., 2017b).

In line with this contemporary view of validity as a unified concept intricately tied to test interpretation and use, this thesis considers validity not just for the primary PIRLS purpose of measuring reading achievement across countries, but also for the specific ways in which PIRLS results are interpreted and used within the South African context.

3.2.2 Conceptualising reliability

Reliability as a foundational component of quality assessment refers principally to the consistency of scores across instances of testing (AERA et al., 2014). As with validity, there have been extensive debates and developments regarding how reliability is understood and evaluated.

Foremost among these is the understanding that reliability is a necessary condition for validity, as interpretations of test scores depend on the assumption that there is consistency of scores across different instances of testing (Kane, 2013).

Early work in the field of psychometrics saw the emergence of mathematical models for estimating reliability in terms of classical test theory indices like test-retest and internal consistency coefficients (Cronbach, 1951; Traub, 1997). In efforts to establish approaches to evaluating the reliability of assessment measures across different instances of test administration, researchers proposed methods for evaluating reliability using model-based procedures that describe an estimate of reliability in terms of standard errors of measurement (Birnbaum, 1969; Cronbach, 1951).

The importance of reliability as a critical psychometric property was more firmly recognised later, when large-scale educational testing became increasingly popular (Brennan, 2001). Psychometricians then began to estimate reliability conditional on population mean and student ability (Lord, 1980; Rasch, 1993). While aspects like test-retest reliability and internal consistency remain important, particularly when considering trend progress across cycles of assessment, a critical issue for evaluating reliability is whether the test can precisely measure proficiency on the construct across the full ability range of all participating students.

The increased application of item response theory (IRT) to ILSAs has seen the emergence of IRT-based methods for estimating reliability that provide an indication of the precision with which item-level performance can be used to estimate the overall ability for each student (AERA et al., 2014). Following work by Mislevy (1991) on randomisation-based inference from complex samples and Rubin (1987) on multiple imputation for missing data, the plausible value approach emerged as a method of estimating scores for students across varied ability levels and from different contextual backgrounds. This approach allows for the use of a large number of items and applies conditioning variables from student background data, both of which help to reduce the variation across estimated scores and improve the precision of the scores estimated (Rutkowski et al., 2013; von Davier &

Sinharay, 2014). Plausible value estimation methods have been generally adopted in ILSAs and are used to estimate reading scores in PIRLS (Martin, Mullis, & Hooper, 2017).

Although this is a widely used method to improve reliability of ILSAs, achieving precise measurement is highly dependent on the overlap between student abilities and item difficulty values across all levels of the construct (von Davier et al., 2020). This is a particular challenge for countries whose students are at the lower end of the proficiency scale as tests that are not well-targeted to low-performing students can result in reduced reliability (Foy et al., 2010; Rutkowski et al., 2019). When item difficulties greatly exceed student abilities, a floor effect occurs which results in extremely imprecise ability estimates and sizable measurement error (Gustafsson & Barakat, 2023; Haertel & Herman, 2005; Rutkowski et al., 2019). This critical issue of adequate measurement precision across all ability levels remains an ongoing psychometric challenge as, even if overall reliability coefficients appear acceptable, measurement precision may still vary considerably for students at the lower end of the distribution (Foy et al., 2010; Foy & Yin, 2017). The increased potential for measurement error poses serious threats to the validity of cross-country comparisons, as well as the validity and reliability of score comparisons for the same country over time (Gustafsson, 2019a).

While concerns have already been raised about the reliability of the consistency of the PIRLS trend results across cycles in South Africa (discussed in Section 2.4), a pressing issue for this thesis results from the threats to the reliability and precision of the 2016 PIRLS achievement measures due to targeting concerns. In line with this modern conceptualisation of reliability as not just consistency, but also the precision with which a test can reproduce stable ability estimates across the full proficiency distribution, this thesis considers reliability from the perspective of both consistent and precise measurement across the ability distribution.

3.2.3 Conceptualising fairness

The concept of fairness is far-reaching, emerging from a rich historical foundation of political, moral and ethical philosophy. As such, it has been the centre of extensive debate around ideas of equity, justice, and human rights. The issue of fairness, therefore, holds special significance in the field of educational assessment which aims to provide information about student traits. Added to this, educational assessments are often used for high-stakes decisions, such as selection for higher education (Crooks et al., 2010; Kunnan, 2008, 2015). However, high-stakes outcomes are not the only contexts in which fairness needs to be prioritised. Fairness is fundamental to validity, and theorists in the field have long recognised that tests must be designed and implemented in a way that is fair to ensure valid inferences can be drawn from the results (Cronbach, 1971; Guion, 1977; Messick, 1980, 1989).

Research into fairness in educational assessment has delved into the consequences of an unfair test by describing issues such as 'construct-irrelevant variance', where factors irrelevant to the construct being measured influence scores (Messick, 1989); and 'differential validity', where the test may not measure the same construct across different subgroups (Cole & Moss, 1989). The presence of construct-irrelevant variance can undermine fairness by advantaging or disadvantaging certain students based on factors unrelated to their true ability on the construct. Similarly, if a test exhibits differential validity, it raises questions about whether the same inferences can be drawn from the scores. Issues such as these raise concerns about the extent to which fair and equal opportunities may be compromised for students from different groups, which has consequences for the validity of the test.

As an understanding of the consequences of unfair tests has grown, so too have the range and prominence of fairness considerations in test design, development, and score interpretation (Clauser & Bunch, 2021; Dorans & Cook, 2016; Nisbet, 2019). This growth in the prominence of fairness considerations can be seen in the evolution of the *Standards*; while earlier editions included references to fairness, the 2014 update allocated an entire early chapter to the concept. This chapter stresses that the concept of fairness in testing is "an overriding, foundational concern" and that if fairness is compromised on the basis of any individual or group characteristic, then it could interfere with the validity of the interpretation of test scores (AERA et al., 2014, p. 49). However, despite the recognition of fairness as a critical consideration in assessment, establishing consensus on its conceptualisation, as well as the practical implementation of fairness in testing situations remains challenging.

Contemporary perspectives recognise that fairness is a complex, multidimensional concept that requires considering equity at all stages of the testing process (Nisbet & Shaw, 2019). Researchers highlight an array of interrelated considerations essential for fairness including setting appropriate expectations and standards that avoid cultural or linguistic bias (e.g., Faulkner-Bond & Sireci, 2015; Ortiz et al., 2018; Reynolds et al., 2021; Shaw & Imam, 2013; van de Vijver & Tanzer, 2004); ensuring fair administration procedures that provide equal opportunities for diverse groups (e.g., Andringa & Godfroid, 2020; Oliveri & von Davier, 2016; Zhao, 2018); ensuring the absence of measurement bias across groups (e.g., Penfield & Camilli, 2006; Reynolds et al., 2021); and ensuring that all students have an equal opportunity to demonstrate their ability on the construct being measured (e.g., De Backer et al., 2019; Jacob & Rothstein, 2016; Ortiz et al., 2018; Ritzhaupt et al., 2013).

In addition to these considerations, fairness is also noted as essential for the valid interpretations and use of test scores (e.g., Crooks et al., 2010; Ercikan & Allalouf, 2016; Kane, 2013). If a test suffers from issues of unfairness, such as construct underrepresentation or construct-irrelevant variance, it threatens the legitimacy of any inferences made from the test scores about the knowledge, skills or abilities it aims to measure. This is of particular concern in ILSAs, where ensuring adequate construct representation, and minimising the risk of construct-irrelevant variance, across a wide range of countries, contexts, languages, and cultures is challenging.

Moreover, fairness considerations extend beyond just the technical quality of the test itself to the social consequences of testing. An unfair test can have adverse impacts on individual students, cultivating feelings of frustration, anxiety and impacting students' academic self-concept (e.g., Lantolf & Poehner, 2016; Messick, 1989; Roever & McNamara, 2006). The consequences of impacts such as low academic self-concept can further undermine motivation and engagement, and subsequently hinder performance on the test for some students (Bieleke et al., 2023; Eklöf, 2010). Moreover, these negative testing experiences can have long-term impacts on students' academic performance more generally, hindering their confidence in reading, and even potentially perpetuating inequality gaps (Chapman et al., 2000; Fulmer & Tulis, 2013). Given these considerable consequences to the student, ensuring fairness in ILSAs requires careful attention and prioritisation.

Nisbet and Shaw (2019) raise another important concern when considering fairness issues in educational assessment. They note that while there is general agreement that fairness relates to the consequences for the students taking the test, most conceptualisations include only a limited discussion of considerations for other stakeholders that may be impacted by test outcomes and the uses of test results. Although ILSAs typically do not have direct consequences on progress or opportunities for participating students, the potential for ILSAs to have an influence on educational policy more broadly is an important consideration.

The role and influence of international large-scale assessment in educational policy

A primary aim of ILSAs is to provide countries with valuable data on student learning outcomes that can help identify strengths and weaknesses in their educational systems (Mullis et al., 2015a; OECD, 2023). Assessments like PIRLS are therefore intended to offer policymakers with an opportunity to monitor progress over time, as well as the ability to benchmark their performance against different systemic approaches (Breakspear, 2012). In some countries ILSAs are viewed as opportunities to derive valuable lessons from comparing their educational systems to others at a point in time and longitudinally, while others see participation as a way to increase their global visibility and influence (Addey et al., 2017; Steiner-Khamsi & Waldow, 2018). In light of this, a number of researchers have looked at the extent to which ILSA results have influenced educational policy discussions, interventions, and reforms in various contexts (e.g., Baird et al., 2011; Crato, 2021; Hopfenbeck et al., 2018; Lewis & Lingard, 2015; Ozga, 2012; Steiner-Khamsi & Waldow, 2018; Stiff et al., 2023).

In many cases, ILSA findings are used to provide evidence-based information about the progress and relative standing of different educational systems (Breakspear, 2012; Ehren, 2021; Gorur & Wu, 2015; Klieme, 2020). This information is often leveraged to reinforce policy priorities that promote fairness and equity (Volante et al., 2022). For example, in the absence of national evaluations in Spain, PISA results have been used to support policies that promote equity across regions (Gomendio, 2021); while in Australia they have been used to advocate for increased funding equity for disadvantaged schools (Thomson, 2021). Similarly, in other contexts (including South Africa) ILSA results have been used to hold governments accountable for addressing educational priorities and areas of concern (Andersen, 2017; Baroutsis & Lingard, 2017; Lewis & Lingard, 2015; Maynier, 2023; van Staden & Gustafsson, 2022; Verger et al., 2019).

However, although ILSAs have the potential to support efforts towards ensuring educational equity, a number of investigations have also revealed the substantial risks that can be associated with using ILSA data for policy-decisions. Concerns have been raised that warn against oversimplifying the complex data (Araújo et al., 2017; Johansson, 2016; Klieme, 2020; Ozga, 2012); inability to account for contextual nuance across diverse systems (Addey et al., 2020; Araújo et al., 2017; Gomendio, 2023; Oliveri et al., 2015); reservations about the reliability of results based on the methodological choices made (Gomendio, 2021; Gustafsson & Barakat, 2023; Klieme, 2020); and placing too much emphasis on league-table standings at the expense of other important educational goals (Addey et al., 2017; Gorur & Wu, 2015; Klemenčič & Mirazchiyski, 2018; Thomson, 2021). There are also concerns about perpetuating inequalities if ILSA data is interpreted unfairly leading to policies that may disproportionately disadvantage certain groups. Added to this, because the majority of participating countries are high-income or developed countries, concerns have also been raised about the fairness and relevance of test content and findings specially for developing countries (Ahmed et al., 2022; Howie, 2012; Kanjee, 2018; Wagner & Castillo, 2014).

As discussed in Chapter 2, PIRLS has high status in the South African context and has already been seen to influence both educational policy as well as the focus areas selected for targeted reforms and interventions. The potential for an unfair test to undermine policy initiatives aimed at improving educational outcomes, is therefore a particularly pertinent issue in the context of PIRLS in South Africa. As such, this thesis conceptualises fairness as a multidimensional and overarching concern for all educational stakeholders; including individual students taking the test, as well as all students and other stakeholders that may be impacted by decisions made related to educational policy and intervention priorities. This means that a number of threats to fairness need to be considered, including comparability of the instruments across language groups as well as the validity and fairness of the way in which PIRLS data and results are used in South Africa.

3.3 Ensuring quality for international large-scale assessments

Ensuring valid, reliable, and fair assessments is a paramount concern for organisations that develop ILSAs. This section provides an overview of some of the common steps and procedures undertaken by test developers to ensure high quality instruments. These procedures are guided by international principles and practices such as those outlined in the *Standards* (AERA et al., 2014) and the International Test Commission's (ITC) *Guidelines* (ITC, 2001, 2005, 2018).

The most recent edition of the *Standards* outlines comprehensive criteria to ensure validity, reliability, and fairness in large-scale assessments like PIRLS (AERA et al., 2014). The 25 standards are organised across three themes: establishing intended uses and interpretations; issues regarding samples and settings used in validation; and specific forms of validity evidence. The ITC also provides *Guidelines* for ensuring quality in international testing that cover issues related to (but not limited to) test use, test design, translation and adaptation of tests, quality control in scoring and analysis, and test security (Bartram & Hambleton, 2016; ITC, 2001, 2005, 2018). Both the *Standards* and *Guidelines* underscore that validity encompasses more than just the construct measured by a test; recommending that interpretations and uses of test scores must also be validated within the given context and for the intended purposes.

Typical quality assurance steps include careful development and updating of assessment frameworks, expert review of items for quality and lack of bias, field testing for statistical evaluation of item characteristics, standardised administration procedures, advanced psychometric scaling, and statistical analyses to evaluate validity, reliability, fairness, and measurement invariance across countries and subgroups (AERA et al., 2014; Rutkowski et al., 2013). As the majority of ILSAs are administered in multiple language versions, ensuring that assessments measure the same construct in different languages is one of the fundamental considerations in designing ILSAs. For tests administered in multiple languages, rigorous translation and adaptation protocols are followed based on guidelines to ensure linguistic and cultural appropriateness (Ebbs & Wry, 2017; Hambleton & Lee, 2013; OECD, 2019). However, establishing comparability can be challenging as the constructs being measured are often complex, and the groups taking the test are wide-ranging. These challenges, together with the increased use of adapted language versions of ILSAs, have led to the development of methods, standards, and guidelines to support the design, translation and administration of the different versions.

The predominant procedures involved in adapting assessments across languages come from recommendations in Chapters 3 and 4 of the *Standards* (AERA et al., 2014); as well as in the *ITC Guidelines on test adaptation* (2017) and the *ITC Guidelines for the large-scale assessment of linguistically and culturally diverse populations* (2019). These provide detailed procedures to ensure appropriate translation protocols are followed to produce assessments with strong psychometric properties suitable for cross-language and cross-cultural comparisons. As a result, the procedures followed in test adaptation for most ILSAs strive to consider the educational, linguistic, cultural, and psychometric properties that may affect different versions of a test (Bartram & Hambleton, 2016; De Boeck & Elosua Oliden, 2016; Rutkowski et al., 2013). The IEA, for instance, follows a rigorous translation and adaptation procedure for PIRLS which is informed by the *Standards* and *Guidelines* as well as other existing literature on international assessment adaptation. These procedures are carefully outlined in the methods and procedures chapter of their assessment framework (Ebbs & Wry, 2017).

Despite stringent protocols, achieving equivalence between language versions is challenging and issues of cultural and linguistic fairness, sociolinguistic dynamics, accessibility, and valid measurement of constructs across versions remain central concerns in reading assessments (Arffman, 2010; Ercikan & Por, 2020; Everatt et al., 2004; Goodrich & Ercikan, 2019; Grisay et al., 2007; Hambleton & Lee, 2013; He et al., 2018; Oliveri et al., 2015). There are ongoing debates regarding whether various ILSAs do indeed effectively (and fairly) achieve their objectives. Typically these debates emphasise concerns around the procedural choices made during the design stages (e.g., Kreiner & Christensen, 2014; von Davier et al., 2019); technical issues of comparability in the analysis (e.g., Elosua Oliden & Mujika Lizaso, 2013; Ercikan & Koh, 2005; Ercikan & Por, 2020; Hambleton & Kanjee, 1995; Huang et al., 2016); as well as concerns about whether it is conceptually feasible to measure the same construct in different contexts fairly and accurately (e.g., El Masri et al., 2016; Maul et al., 2019; McGrane & Maul, 2020). Added to this, it is widely held that the factors that affect test comparability go beyond just an evaluation of form and properties; ensuring that language versions represent the same level of difficulty for each group is fundamental to fair comparisons (Arffman, 2010; Hambleton et al., 2004; Hambleton & Lee, 2013). Concerns about the burden placed on students, as well as the potential consequences of test anxiety are also prevalent, particularly in cases where students in specific groups may face differential levels of difficulty from the same test (DeMars & Wise, 2010; Eklöf, 2010; Wise et al., 2020). The next section details the key threats that critiques such as these pose to upholding validity, reliability, and fairness in the context of PIRLS in South Africa.

3.4 Key areas of concern for PIRLS in South Africa

This review has covered fundamental concepts in ensuring the quality of assessments by discussing important conceptual, practical, and methodological challenges inherent in assessing reading achievement across languages and contexts, as well as the steps taken by test developers to mitigate these issues. Through this lens, the sections that follow highlight specific threats to valid, reliable, and fair assessment of reading comprehension in the South African context. This thesis focuses on three central threats to guide the empirical investigation of PIRLS: contextual relevance, wide gaps in achievement, and comparability across language versions. These threats are particularly pertinent given South Africa's diverse linguistic and cultural landscape, coupled with persistent inequalities in educational outcomes, which add complexity to ensuring valid use and interpretations of PIRLS results. A comprehensive review of the literature related to each threat is provided within the respective papers in Chapters 5, 6, and 7; the following sections offer a concise overview to establish the foundation and rationale for subsequent analyses.

3.4.1 Threat 1 – Contextual relevance

One of the dominant criticisms of ILSAs like PIRLS is their ability to adequately attend to specific contextual considerations in each of the contexts in which they are administered and used (Addey et al., 2020; Gomendio, 2023; Klemenčič & Mirazchiyski, 2018). When an assessment lacks contextual relevance, it can pose a significant threat to the validity and fairness of the inferences drawn from the test scores (Lin et al., 2014; Oliveri et al., 2018). To address these concerns and in an effort to mitigate some of the challenges associated with the diversity across contexts, PIRLS includes a number of contextual questionnaires that aim to capture data specific to each context (Hooper et al., 2015). In addition to this, participating countries are able to make adaptations to the items included in these questionnaires to ensure they capture data relevant to the national context (Foy, 2018b).

The importance of contextual relevance is heightened when considering the administration of PIRLS in the highly diverse and unequal educational landscape of South Africa. Despite its status as a key assessment that is used to inform educational policy in the country, questions have been raised about the validity and comparability of PIRLS results given the country's complex linguistic, cultural, and socioeconomic dynamics (Mtsatse & van Staden, 2021; Roux et al., 2022; van Staden et al., 2019; Zimmerman & Smit, 2016). One of the areas highlighted in discussions of PIRLS results in South Africa is the wide gaps in performance seen across language groups (Spaull & Pretorius, 2022). However, due to the enduring effects of apartheid policies and the stark divides in school quality, simply considering differences in group averages may mask significant disparities across different student groups facing vastly different contextual experiences. As a result, the assessment may fail to accurately capture students' true reading abilities, leading to misleading interpretations about achievement gaps across different demographic groups. This is particularly relevant as socioeconomic status and educational opportunities can further exacerbate academic disparities leading to longer-term implications if the evidence used to inform educational decisions is invalid.

Gaining nuanced insights into the specific factors influencing reading achievement within South Africa's intricate educational context is therefore essential for accurately interpreting PIRLS data and informing targeted, contextually relevant interventions. To address this threat, Paper 1 aims to gain a more detailed understanding of the reading achievement of South African students that goes beyond surface level comparisons of average achievement. This paper conducts a thorough analysis exploring how key contextual factors relevant to South Africa – including language background, medium of instruction, socioeconomic disparities, and variations in school quality – relate to student performance in PIRLS 2016. It addresses three research questions:

1.1 To what extent do contextual factors impact students' overall PIRLS 2016 reading achievement in South Africa?

1.2 To what extent do factors related to the use of the test language at home impact students' overall PIRLS 2016 reading achievement in South Africa?

1.3 To what extent does the test language impact students' overall PIRLS 2016 reading achievement in South Africa?

By considering the nuanced relationships between language, context, and reading achievement, Paper 1 provides a foundational grounding for subsequent investigations of the PIRLS instrument undertaken in Paper 2 and Paper 3 by providing critical insights into how the South African context shapes reading achievement in PIRLS. This contextualised analysis enables more detailed interpretations of the results and highlights areas where increased contextual relevance may be needed for fair and equitable assessment across South Africa's diverse student population.

3.4.2 Threat 2 – Wide gaps in achievement

One of the major threats to the validity, reliability, and fairness of ILSAs emerges from the wide variations in achievement levels across the diverse range of participating countries. As the number of countries and systems participating in ILSAs continues to grow, so does the challenge of developing instruments and methods that can reliably and validly measure the construct across the range of abilities. When considering the widening participation in PIRLS in particular, threats to validity include factors such as curriculum misalignment as more countries with more varied curricula participate, as well as accuracy when scaling the scoring distributions across an extreme range of achievement levels (Kieffer et al., 2009; Lin et al., 2014).

The complexity of ILSAs and the diversity of the contexts in which they are applied means that they are highly sensitive to the methodologies they apply, and the validity and comparability of the results is dependent on those methodologies (Ercikan, 1998). If the methodological model does not adequately fit the data from certain contexts, assumptions of the model may be violated and concerns are then raised about the reliability and validity of the scores estimated for students in those contexts.

There are two major threats to valid measurement related to the widening achievement ranges relevant to PIRLS in South Africa. Firstly, the large gap between the highest and lowest performing countries, with the highest performing countries achieving averages above 570 points, and South Africa, the lowest performing country, achieving an average of 320 points. These wide disparities in achievement across countries raise concerns about the reliability of the scores estimated, particularly for the low performing groups (Foy et al., 2010). Additionally, as the reliability of the estimated scores is called into question, so too is the validity of any inferences that can be made from the results (Oliveri et al., 2018).

The second notable concern is the substantial gaps in PIRLS performance seen across groups within South Africa. In every PIRLS cycle, the spread between the top and bottom performers within South Africa has been wider than the achievement distributions of all other participating countries. These wide gaps are reflective of the extreme inequality seen in the country, and the gaps between language groups have been emphasised as a major area of concern and attention for educational policy as well as targeted interventions. Another challenge regarding these internal gaps, is the extent to which they compound the targeting concerns already present due to South Africa's low overall achievement internationally.

Appropriately targeting the difficulty of items in a test to the abilities of the students taking the test is a fundamental requirement of quality assessment. Poor targeting is of particular concern for low-performing countries because, when students at the lower end of the achievement continuum encounter items that are too difficult for them, it can result in flooring effects that reduce measurement precision and reliability for those groups (Gustafsson & Barakat, 2023; Rutkowski et al., 2019). Added to this, poor targeting also introduces construct-irrelevant variance as students may guess answers because they lack the baseline ability assumed by the test (Rutkowski & Rutkowski, 2021). This threatens the comparability of the scores while also failing to provide those students with the opportunity to adequately demonstrate their skills. Furthermore, this can negatively affect the motivation, self-concept, and self-esteem of students from low-performing groups and lead to concerns about the fairness and equity (Shen & Tam, 2008; Yoshino, 2012; Zhao, 2016).

In light of these threats, Paper 2 aims to evaluate the extent to which the models used by PIRLS exhibit adequate fit to the data from groups within South Africa, as well as the extent to which the assessment is appropriately targeted to the abilities of students. Focusing on three language groups – the two highest performing and the lowest performing – it considers the following research questions:

2.1 To what extent does the model used by PIRLS 2016 fit students' data from the Afrikaans, English, and Sepedi groups in South Africa?

2.2 To what extent is the difficulty level of items in PIRLS Literacy 2016 suitably targeted to the ability of students in the Afrikaans, English, and Sepedi groups in South Africa?

By addressing these research questions, Paper 2 aims to provide insight into potential methodological threats that could undermine valid and comparable measurement for South African students in PIRLS.

3.4.3 Threat 3 – Comparability across language versions

A major threat to the validity and fairness of score comparisons in ILSAs like PIRLS is ensuring the comparability of construct measurement and difficulty across the multiple language versions. Given that differences in difficulty or construct realisation across versions can threaten score comparability, and in order to ensure fair opportunities for all students to demonstrate their skills, extensive procedures and quality assurance checks are followed when adapting ILSAs into different language versions (Bartram & Hambleton, 2016; Ercikan & Allalouf, 2016; ITC, 2017; Oliveri et al., 2015).

Adapting an assessment from a source language to multiple target languages while maintaining equivalent construct representation and psychometric properties across versions is a considerable undertaking. The extent of the potential difficulties with establishing equivalence across languages have led some researchers to argue that true equivalence may not be possible, particularly where languages vary substantially in form and syntax (Arffman, 2010; El Masri et al., 2016; Grisay et al., 2007; Pym, 2009). This is particularly pertinent in cases such as PIRLS where language forms an integral part of the construct of reading comprehension. Instead of focusing on equivalence, therefore, this thesis considers *comparability* across versions as a key component of ensuring fair and valid cross-language assessments.

The process of translating all the required instruments while maintaining the comparability across languages is one of the most difficult aspects of implementing PIRLS effectively in South

Africa. In total, 18 test instruments and two questionnaires needed to be translated from English into 10 languages, meaning that more than 200 different versions of the instruments were created for the South African context (Howie, Combrinck, Roux, Tshele, et al., 2017). The procedures involved first adapting the US English source versions of the various instruments to UK English as well as adapting necessary aspects (such as terminology) for the South African context before sending back to the IEA for verification. Once verified, the adapted English versions were then sent to a certified translation company in South Africa (Dilicom) to be translated into the remaining 10 official languages (Howie, Combrinck, Tshele, et al., 2017). These translations were then sent back to IEA for verification and the comments were addressed by the local translation teams. The final process of translation verification involved consultation with local language specialists as well as a team of primary school teachers in all languages to assist with ensuring that the language used was appropriate for primary school students. Despite the extensive efforts involved, concerns about the extent to which comparability is realised across these language versions continue to emerge (Mtsatse & van Staden, 2021; Roux et al., 2018, 2022).

Another notable concern regarding the comparability of PIRLS in the South African context is the wide discrepancies in overall achievement across language groups. The extensive differences seen in the achievement between students who took the test in Afrikaans or English, versus those who took the test in an African language, suggest that there may be cause for concern regarding the comparability of the construct and difficulty across versions within the country. While various psychometric methods exist for evaluating cross-language comparability (Bauer, 2023; Cohen & Bolt, 2005; Penfield & Camilli, 2006; Robin et al., 2003; Sireci & Rios, 2013); much of the ILSA research centres around cross-country comparability, and equal scrutiny is not applied at a within-country level (e.g., Atar et al., 2023; Gökçe et al., 2021; Grisay et al., 2009). However, evaluating within-country comparability is crucial in multilingual contexts like South Africa to ensure valid score interpretations and fair testing for all students.

To investigate this key threat, the final paper in this thesis examines the comparability of the Afrikaans, English, and Sepedi versions of PIRLS 2016. Given the structure of the PIRLS instruments – reading passages with associated items – three research questions are addressed in Paper 3:

3.1 To what extent are the Afrikaans, English, and Sepedi versions of PIRLS Literacy 2016 comparable at an item level?

3.2 To what extent does the comparability of PIRLS Literacy 2016 across these languages at an item level impact on the comparability of the assessment at a passage level?

3.3 To what extent does the comparability of PIRLS Literacy 2016 across these languages at item and passage level impact on the comparability of the assessment at test level?

By investigating comparability at the individual item level, at passage level, and at the overall test level, this analysis aims to provide comprehensive insights into the extent to which any comparability issues found may impact the fair and valid interpretations of PIRLS results across language groups in South Africa.

3.4.4 Summary

The three threats outlined here represent key validity and comparability challenges that need to be carefully examined to ensure the appropriate use and interpretation of PIRLS results in South Africa. Although each of these threats are presented and described separately in this chapter, it is important to note that there are substantial overlaps between them in the diverse South African educational landscape. As described in more detail in Chapter 2, historically unequal education policies are likely to influence the wide achievement gaps seen, as well as potential issues of contextual relevance and cross-language comparability. It is therefore essential that the interdependencies between these threats are considered carefully in the context of South Africa to ensure a holistic approach throughout the investigations undertaken in this thesis. The next chapter outlines the specific methods and analytic techniques employed across each paper to evaluate these issues as well as details about the dataset used for the investigation.

CHAPTER 4 METHODOLOGY

This thesis investigates PIRLS 2016 in South Africa by exploring both the impact of the broader context on student performance, and by evaluating the psychometric properties of the PIRLS instrument. The first paper addresses the contextual influences on PIRLS 2016 results in South Africa, with a specific focus on understanding the relationship between reading achievement and test language in South Africa through the application of multiple linear regression modelling.

After establishing a picture of the contextual influences on PIRLS reading achievement in Paper 1, the investigation turns to the test instrument. The psychometric properties of the instrument are evaluated in Papers 2 and 3, where fundamental questions of the fit, validity and comparability of PIRLS Literacy 2016 in South Africa are considered from within an IRT framework. This chapter provides an overview of the data and methods used to conduct the analyses for both the contextual exploration and the psychometric evaluation and serves to compliment the more focused methods sections provided in each of the three papers that form this integrated thesis.

4.1 Data

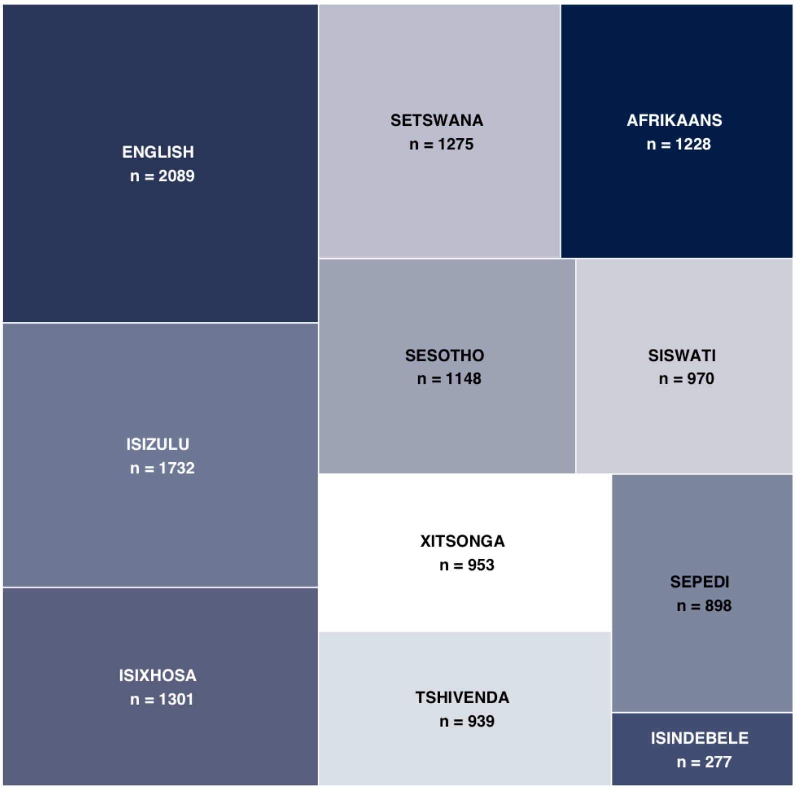
This research uses secondary data from the publicly available PIRLS 2016 international database to answer the research questions in all three papers. Although there are limitations to using secondary data, there are also considerable advantages. The primary advantage is the wealth and scale of information that can be accessed through existing datasets without the time-consuming and often limiting options present when collecting primary data. Projects such as PIRLS adhere to rigorous data collection standards and provide detailed information about the data and methodology that enables secondary analysis on multiple levels. In the case of this research, PIRLS Literacy 2016 has an additional advantage specific to the South African context. Although there have been large-scale reading assessments conducted in all 11 languages previously (see Section 2.3 in Chapter 2 for details), PIRLS is currently the only reading assessment in primary schools conducted in 11 official languages at a national scale.

One of the challenges that can arise with large-scale data is, due to the large and often complex datasets involved, it may take a substantial amount of time to identify the variables required for analysis, understand the detailed components of these variables, and then import, merge, code, and clean the data before analysis can even begin (Smith, 2008). The IEA provides a substantial amount of detail on PIRLS data that is publicly available, as well as guidelines to assist researchers with the complex task of working with such a large data set via their website, documents such as the detailed codebook, and assessment frameworks. Additionally, detailed user guides are available that provide secondary researchers with the tools required to embark on analysis of the dataset rigorously and accurately (Foy, 2018a). Nevertheless, some recoding, transformations and preparation was required before analysis could begin. A description of the data preparation and recoding that took place before analyses could be conducted is included in Appendix B.

The data selected for secondary analysis in this thesis features data from Grade 4 students in South Africa who participated in PIRLS Literacy 2016, and includes information about the students, their participation in PIRLS, their achievement results for PIRLS, as well as individual response data for each of the items across the 12 PIRLS Literacy 2016 reading passages. The data also includes responses to four questionnaires from the students, their primary caregivers at home, their teachers, and principals at the schools they attended. The sections that follow describe the student sample and instrument data used for the analysis in this thesis.

4.1.1 Sample

The PIRLS 2016 national dataset for South Africa is a nationally representative sample of students in Grade 4 from all nine provinces, and 11 language groups. The PIRLS Literacy 2016 national dataset for South Africa (N=12,810) was retrieved from the IEA International Data Repository using the EdSurvey package in R (Bailey et al., 2020). Figure 4-1 shows a treemap of the dataset with blocks portraying the number of students from each language group. The English group had the most students (2,089), followed by isiZulu (1,737), isiXhosa (1,301), Setswana (1,275), Afrikaans (1,228), Sesotho (1,148), siSwati (970), Xitsonga (953), Tshivenda (939), Sepedi (898), and isiNdebele (277).



Source: IEA International Data Repository for PIRLS 2016

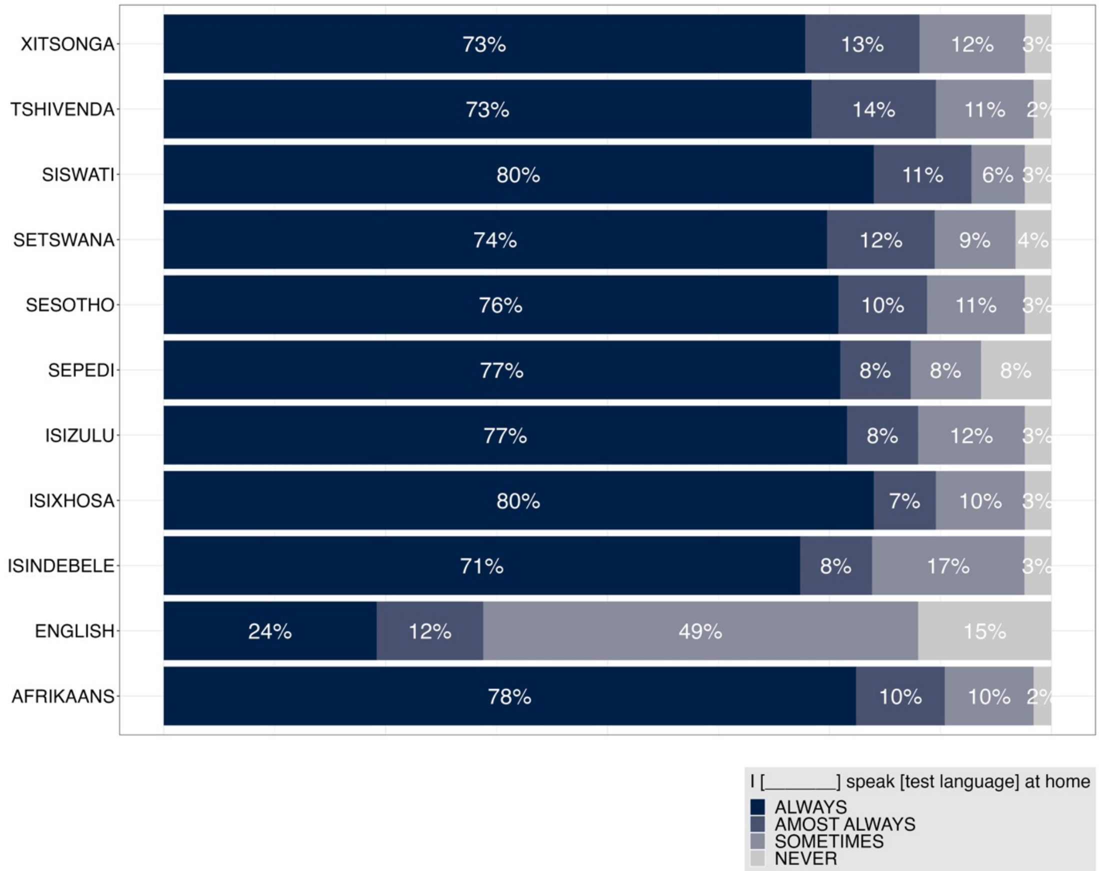
Figure 4-1 Number of students per language group in the PIRLS 2016 South African dataset

While the sample in this study is nationally representative, the size of each language group does not accurately reflect the relative frequency at which these languages are used as a medium of instruction in South African primary schools. In South Africa, over 70% of students are taught in an official African language, yet the PIRLS 2016 sample includes a high number of English classes (and therefore students) compared to the other language groups. This discrepancy is due to implicit stratification in the sampling design of PIRLS. PIRLS uses a two-stage random sampling design to select schools for participation in the international study (Martin, Mullis, & Foy, 2017). In the first stage, a sample of schools is selected in each country and in the second stage, one or two intact classes from a specific grade-level are selected from each of those schools. PIRLS selects whole classes rather than individual students because of their focus on the curricular and instructional experiences related to reading achievement and to ensure minimal disruption to the day-to-day school routine (Martin, Mullis, & Hooper, 2017).

PIRLS 2016 also used stratification during the first stage of the sampling to improve efficiency of the sample design, to enable the application of different sample designs to specific groups of schools, and to ensure proportional representation of specific groups in the sample (Martin, Mullis, & Foy, 2017). In the case of South Africa, the stratification during the first stage included: medium of instruction, province, level of urbanisation, type of school (public or independent), and the socioeconomic categorisation of schools (known as quintiles) (Howie, Combrinck, Roux, Mokoena, et al., 2017). Due to the complex multilingual nature of schools in South Africa, there are also schools with multiple 'language streams', meaning that one school may have two or more options for their media of instruction. In other words, it is possible for some schools to have two or more classes with different media of instruction in a single grade-level (e.g., one for English-speaking students and one for Sepedi-speaking students). It is typical for one of these options to be English, while the other language varies greatly by location (Howie, Combrinck, Tshele, et al., 2017). This means that there are a higher number of English classes, and by extension students in the English group than in other groups in the sample.

All students in this dataset were in Grade 4 and the average age for the sample was 10 years and 7 months. An important consideration is that for a majority of students in South Africa in Grade 4 the medium of instruction in all subjects switches to English or Afrikaans (in most cases from students' home language to English). The challenges associated with the transitions in Grade 4 are discussed widely in research and teaching circles in South Africa (e.g., Broom, 2004; Evans & Nthulana, 2018; Shepherd, 2018; Zimmerman, 2014). Due to the switch in medium of instruction in Grade 4, the PIRLS Literacy study that was conducted with Grade 4s in South Africa in 2016 opted to test students in the language that was their medium of instruction in the previous phase of schooling (i.e., their medium of instruction from Grade 1-3). It is important to note that although the language students were tested in was no longer the medium of instruction in Grade 4, it is still considered a 'home language subject' and is still taught for a minimum of six hours a week as a subject.

The challenges with drawing inferences and comparisons on the basis of language are numerous and one of the primary issues raised is whether or not the test language is the same language students speak at home (Akkakoson, 2013; Howard et al., 2014; Yamashita, 2002). Although this is not one of the primary aspects being investigated in this thesis, due to its relevance and potential influence, information about the rate at which home language is spoken across groups is provided in Figure 4-2. In almost every group a majority of students (73%-80%) spoke the test language as a home language. Only the English group, which is the highest achieving group, had a much lower percentage of students (24%) who always spoke the test language at home.



Source: IEA International Data Repository for PIRLS 2016

Figure 4-2 Frequency of test language use at home by South African students

4.1.2 Instrument data

The PIRLS 2016 data also includes response data and overall results from two main instruments. The reading test which is made up of 12 passages and approximately 15 items for each passage, and the context questionnaires which are made up of four separate background questionnaires (student, home, teacher, and school) each with 10-14 items. Different subsets of the full national sample, and different sets of items from the test and/or context questionnaires are used in each of the three papers to answer the research questions most effectively. A more detailed description of the PIRLS instrument is provided in Appendix A.

4.2 Exploring reading achievement in the South African context

Paper 1 explored the impact of contextual factors on PIRLS Literacy 2016 performance in South Africa with a specific focus on the relationship between test language and overall reading achievement. Multiple linear regression was used to model the relationships between student, home, and school contextual factors and language variables on overall reading achievement. The dependent variable is the set of five plausible values (PVs) representing students' overall PIRLS reading achievement scores. Predictor variables encompassing relevant contextual factors were selected based on demonstrated relationships with reading achievement both internationally (Mullis et al., 2017a), and in the South African context (Howie, Combrinck, Roux, Tshele, et al., 2017). In total, 21 predictor variables across student, home, and school contexts were included, as well as variables related to language use at home and test language. More details about the PVs in PIRLS, as well as the set of predictor variables selected for this analysis are included in Section 5.3 in Paper 1.

Multiple linear regression modelling was selected as it is well-suited for simultaneously examining multiple predictors and their interactions while estimating effects on the dependent variable (Cohen et al., 2013; Grömping, 2012). Furthermore, this approach accounts for potential confounding between contextual variables and allows for interpreting the magnitude of the estimated effects of predictors on reading achievement (Ziglari, 2017). Multiple linear regression modelling also aligns with precedents set in other investigations of contextual influences on PIRLS achievement (e.g., Guzmán-Simón et al., 2020; Taylor & Yu, 2009; van Staden et al., 2019).

Student sampling weights and Jackknife repeated replication techniques were applied in the computation of achievement and contextual questionnaire data to account for the complex sampling design used in PIRLS data collection (Foy, 2018; Rutkowski et al., 2010). The nested data structure was addressed through the use of sampling weights rather than multilevel modelling due to the selection of one whole class per school that forms part of PIRLS sampling design. A more detailed description of the modelling approach taken as well as the analytic steps followed are included in Section 5.3 in Paper 1. Prior to conducting the analysis, regression diagnostics were examined carefully to ensure the assumptions of the model were met sufficiently and these are included in Appendix C-2.

By leveraging the PIRLS contextual data in South Africa and modelling the relationships between contextual factors, language backgrounds, and overall reading achievement, Paper 1 provides important groundwork for understanding differential performance in PIRLS Literacy 2016. The findings pave the way for the in-depth psychometric investigations that follow in Papers 2 and 3.

4.3 Evaluating the psychometric properties of PIRLS Literacy 2016 in South Africa

The next two papers look at the PIRLS Literacy 2016 instrument from a psychometric perspective. Paper 2 considers the fit, targeting, and validity of the test and items for students in South Africa; and Paper 3 considers the comparability of the instrument across language versions at an item, passage, and overall test level.

4.3.1 Selecting language groups

For Paper 2 and Paper 3, a subset of three language groups (Afrikaans, English, and Sepedi) was selected from the larger national dataset used for Paper 1. This subset includes a total of 4,215 students across three groups; the English group had the most students (2,089) followed by Afrikaans (1,228) and then Sepedi (898).

The choice of these three languages is based primarily on the opportunities for comparison offered by the relative performance of these language groups in PIRLS 2016. Students who took the test in Sepedi were the lowest performing group, with only 7% of students able to demonstrate the ability to read at a basic level and no students reaching the international average. Added to this, the findings from Paper 1 revealed that students in the Sepedi group were likely to achieve a substantially lower score than students in the English group even after controlling for contextual factors and the extent to which the test language was spoken at home (see Section 5.5.4 in Paper 1).

Furthermore, although there is a dearth of research into reading in African languages generally (see Mohohlwane, 2019); research that is conducted in the field of educational assessment typically features languages from the Nguni family (e.g., isiZulu and isiXhosa), which are among the most commonly spoken languages in South Africa. Sepedi was therefore selected as one of the focal languages due to the low performance of students from that group relative to English and Afrikaans, the scarcity of research that focuses on reading in African languages, and the paucity of research focusing on languages from the Sotho language family in particular.

4.3.2 Item response theory framework

For the psychometric analyses conducted in both Paper 2 and Paper 3 an IRT framework is used. IRT is the most commonly employed approach for the design, construction, and analysis of large-scale assessments internationally and it is the core framework used for the design, review, and analysis of both the PIRLS 2016 reading test and the contextual questionnaires (Martin, Mullis, & Hooper, 2017). In this study, the IRT framework used by PIRLS is applied to conduct a secondary analysis of PIRLS Literacy 2016 data in South Africa.

This means that the secondary analysis can be applied using a framework and statistical models that account for the properties of the test items and the abilities of the students. However, although these models were used in the assessment design and analysis for PIRLS Literacy 2016, that is not sufficient to justify their application in a secondary analysis such as this. Fundamental questions that need to be asked when choosing a measurement model include whether the model adequately fits the data, and whether the model fits the approach being taken for the investigation (Harris, 1989; van Rijn et al., 2016; Zhao & Hambleton, 2017). Moreover, the low achievement of students in South Africa in PIRLS may pose challenges to the appropriateness of using these models, as the fit and estimation procedures of the models can be influenced by the distribution and range of the scores across participating countries (Rutkowski et al., 2019; Tijmstra et al., 2020). It is therefore necessary to carefully consider the potential impact of low overall achievement on suitability of the models applied as well as the limitations and consequences that may arise from applying models from an international study intended for cross-country investigations, to subsets of national data for within country investigations.

Three distinct models of IRT were used by PIRLS in the design and analysis of their assessment data: a two-parameter logistic (2PL) model, a three-parameter logistic (3PL) model, and a generalised partial credit model (GPCM). The choice of model is based on the item type (i.e., multiple-choice or constructed response) and the scoring (i.e., dichotomous or polytomous) (Martin, Mullis, & Hooper, 2017). The 2PL model incorporates two parameters: item difficulty and item discrimination. It is suitable for dichotomous items (i.e., 1 – correct and 0 – incorrect). The GPCM is designed for polytomous items, where partial credit can be awarded for an item (e.g., 2 – correct, 1 – partially correct, and 0 – incorrect). The GPCM, therefore, includes threshold parameters that for account for varying levels of item difficulty across the different score categories.

The 3PL model extends the 2PL model by including a guessing parameter, which accounts for the possibility of students guessing the correct answer on multiple-choice items. Hence, the 3PL model is typically used for multiple-choice items. In PIRLS 2016 parameters of constructed response items were estimated using a 2PL model for dichotomous items and GPCM for polytomous items, while parameters for multiple-choice items were estimated using the 3PL model.

For this analysis, consistent with the approach used by PIRLS 2016, a 2PL model was used to estimate parameters for dichotomous items, and a GPCM was used for polytomous items. However, although the 3PL model was used for PIRLS 2016, the subset of data analysed in Paper 2 and Paper 3 does not provide sufficient sample size for accurate and stable estimation of the third parameter, the guessing parameter, and so the 2PL model was used to estimate parameters for the multiple-choice items (Embretson & Reise, 2013). This allows for more stable and accurate estimation of student and item difficulty parameters thereby reducing the potential for item and model misfit (van Rijn et al., 2016). Detailed descriptions of the 2PL model and GPCM, the model assumptions, as well as the equations and parameters that represent each model are presented in Section 6.3 in Paper 2. The sections that follow give a brief overview of the approaches to evaluating the fit and comparability of PIRLS 2016 in South Africa undertaken in Paper 2 and Paper 3 respectively.

4.3.3 Evaluating fit

The low overall performance of South African students in PIRLS Literacy 2016, together with the relative differences in achievement across language groups, means that the assumption that the test is fit for the purpose of assessing reading in South Africa is not guaranteed. To thoroughly interrogate the validity of PIRLS Literacy 2016 for these students, therefore, issues of model and item fit are considered carefully in Paper 2. This is a particularly important step for secondary analyses such as this because a smaller subset of student response data (compared to the primary international set) is being analysed which has the potential to introduce issues of fit (Nye et al., 2019).

The evaluation of fit in Paper 2 considers three primary aspects: the fit of the models used, potential violations to the underlying assumptions of unidimensionality and local independence that characterise IRT models, and the alignment between the ability of students and the difficulty of the items. To answer research question 2.1 (To what extent does the model used by PIRLS 2016 fit students' data from the Afrikaans, English, and Sepedi groups in South Africa?), model fit was considered using comparisons of Akaike Information Criterion (AIC) and the Bayesian Information Criteria (BIC), as well as a likelihood ratio test (Lumley & Scott, 2015). The assumption of unidimensionality of the items was analysed using an exploratory factor analysis technique known as item factor analysis (IFA) (Osteen, 2010); and the assumption of local independence was investigated using Yen's Q_3 statistic, which gives a measure of the correlation between item residuals (Yen, 1984).

To answer research question 2.2 (To what extent is the difficulty level of items in PIRLS Literacy 2016 suitably targeted to the ability of students in the Afrikaans, English, and Sepedi groups in South Africa?), the alignment between student ability and item difficulty was inspected graphically by considering the expected item response functions (IRF) for each item alongside the observed responses. Additionally, the alignment between the distribution of all items in terms of their difficulty and the distribution of all students in terms of their ability was inspected using a Wright Map, which presents the distribution of ability for the students taking the test and the distribution of the difficulty of the items in the test along the same scale (Wright & Stone, 1979). A more detailed description of the criteria used to evaluate fit, as well as the analytic procedures employed are described in full in Section 6.4 in Paper 2.

4.3.4 Evaluating comparability

The final paper in this thesis involves a psychometric evaluation of the comparability of items across language versions, and the impact of any comparability issues found at the passage, and overall test level. Consistent with the IRT approach used for Paper 2, a combination of the 2PL model and GPCM is used to estimate parameters for Paper 3.

Differential item functioning (DIF) analysis is conducted using an IRT-based likelihood ratio testing framework, IRT-LR, to detect DIF at an item level and answer research question 3.1 (To what extent are the Afrikaans, English, and Sepedi versions of PIRLS Literacy 2016 comparable at an item level?) (Thissen et al., 1993). IRT-LR is selected as a method for identifying DIF in this study because it allows flexibility in the choice of IRT model applied (e.g., 2PL model) and it can account for polytomous item response models (e.g., GPCM) effectively (Chen & Hwu, 2018). Comparisons were conducted pairwise with English as the reference group and Afrikaans or Sepedi as the focal group in each pair. As detected DIF can be negligible in terms of effect size, DIF magnitude at the item level was quantified using the approach suggested by Meade (2010), who recommends considering multiple measures of effect size to understand any potential DIF effects fully. More details about the effect size indices used, as well as the approach to understand the nature and direction of DIF is described in Section 7.4 in Paper 3.

Once DIF is identified, two key questions typically arise – ‘what causes DIF to occur?’ and ‘what is the overall impact of DIF found?’. The first question is commonly investigated using an explanatory approach, where language experts, translators, and content specialists review items to identify potential sources of DIF based on features, content, topics, or cultural references that could disadvantage a group (e.g., Ercikan et al., 2010; Huang et al., 2016; Roth et al., 2013). This approach is valuable for understanding the root causes of DIF to inform future test and item development. However, it's important to note that in many cases, it can be challenging to conclusively explain DIF as it can arise from a multitude of factors, some of which may not be easily identifiable. A key consideration therefore remains regarding the overall impact of identified DIF on test score comparability, as DIF effects may accumulate across items and result in systemic differences in how the test functions for particular groups (Chalmers, 2022; DeMars, 2011; Heinz et al., 2022).

The evaluative approach to understanding DIF is commonly referred to as differential test functioning (DTF). DTF originated in pioneering work by Wainer (1993) and Raju et al. (1995) who proposed statistical methods to assess overall test bias resulting from DIF items favouring one group. DTF may occur because a large number of items in a test contain DIF for a specific group or in situations where small differences in item functioning in the same direction for one group may aggregate and result in the test functioning differently for that group (Chalmers et al., 2016; Meade & Wright, 2012; Temel, 2023). DTF approaches, therefore, involve quantifying the cumulative effect of DIF found at the item level (in terms of effect size and direction) to determine whether the item-level DIF results in meaningful score differences between groups.

While different approaches for identifying DIF are widespread, the same cannot be said of DTF. DTF is applied less frequently to investigations of test comparability across groups and as a result there are not as many methods available for analysing DTF (Chalmers, 2018). This should not be construed as devaluing the importance of investigating DTF. In many cases, particularly in secondary analyses such as this, understanding the way in which an instrument functions across groups at the test level is of primary importance to understanding the comparability of test versions. Moreover, it is a particularly important consideration for PIRLS in South Africa as differences in overall scores across groups have been used to justify and inform interventions and policy reforms (see Section 2.4.2). It is therefore critical to understand whether DIF identified at an item level has an impact on the comparability of overall scores across groups.

Approaches to identifying DTF using an IRT framework include Raju's DFIT statistics (Raju et al., 1995), compensatory SIBTEST methods (Shealy & Stout, 1993), and DTF statistics (Chalmers et al., 2016; 2018). This study has selected to use the differential response functioning (DRF) framework proposed by Chalmers (2018) as it is able to account for variability in item parameter estimates as well as correcting for differences that may be the result of variation in ability distributions, which is a feature of the groups being compared in this analysis (Chalmers, 2022; Temel, 2023). Furthermore, the DRF statistics used are able to describe the cancellation effects of non-uniform DIF both in terms of the direction of impact (i.e., which group, if any, is disadvantaged overall), and the absolute impact (i.e., what is the overall unsigned difference across groups) (Bauer, 2023; Finch & French, 2023; Teresi et al., 2021). A further benefit of using these statistics is that they are presented in the metric of the total score which means they are straightforward to interpret in terms of score comparability (Gnambs & Lenhard, 2023; Schöfl et al., 2023).

The impact of DIF in this study is considered at both a passage and test level. Passage level considerations are particularly important in PIRLS as the matrix block administration design means that students respond to different subsets of items, potentially facing inconsistent levels of difficulty. Two key DRF statistics are considered to answer research questions 3.2 (To what extent does the comparability of PIRLS Literacy 2016 across these languages at an item level impact on the comparability of the assessment at a passage level?) and 3.3 (To what extent does the comparability of PIRLS Literacy 2016 across these languages at item and passage level impact on the comparability of the assessment at test level?). First, signed differential response functioning (sDRF) statistics are considered, which quantify expected score differences between groups favouring particular directions at either passage or test level. Next, unsigned differential response functioning (uDRF) statistics are considered, which describe the absolute deviation of differences in scores at the passage or test level, irrespective of the direction of the differences. More details about the DRF approach undertaken, as well as equations describing the statistics used to quantify DIF impact are presented in Section 7.3.4 in Paper 3.

4.4 Methodological limitations and considerations

As with any approach to research, using secondary data comes with advantages as well as potential limitations and challenges. One of the primary limitations when dealing with secondary data is that researchers do not have input into the research design and framework according to which the data is collected. This can cause limitations for the researcher embarking on secondary analysis as aspects such as the original data collection methods, the sampling approach, the questions asked in the questionnaire, or even which participants are included may not meet the needs of the secondary researcher (Little et al., 2013).

PIRLS 2016 is helpful in this regard as it provides detailed information regarding the frameworks according to which the research was approached, the data collection and research design aspects, the design of the approach and instruments for collecting contextual information, the theoretical framework according to which the construct is conceptualised, as well as details regarding the sampling, data handling, coding, and data analysis aspects. In addition to this, there are individual country reports detailing aspects of the project design, rollout or analysis that may be specific to a particular country.

Nevertheless, a challenge that faces researchers embarking on analysing existing data involves the capacity of data sets to answer the research questions. In some cases, the research question that a researcher may seek to address cannot be tackled comprehensively by the dataset available and a researcher may be forced to reframe their question. This might mean that nuanced areas of interest and possible findings may be overlooked in favour of what is available, thereby losing out on the certain aspects of the freedom of research inquiry.

In the case of this research project, the research questions have been led by the dataset to a certain extent. However, the broad area of interest that initially guided this research – to better understand reading in the multilingual context of South African primary schools – remains central to the primary aims of the study. Some of the considerations that are important to note regarding this investigation include concerns about the scope and relevance of the questionnaire data, and methodological limitations related to the sample size and low overall performance of South African students on the test.

4.4.1 Questionnaire data

The PIRLS questionnaire scales aim to describe contextual factors like socioeconomic status, home resources for learning, parents' education levels, and other relevant information. To facilitate more reliable information across contexts, participant responses to multiple related item sets are derived to produce one scaled variable that relates to a single construct (e.g., confidence in reading) (Martin, Mullis, & Hooper, 2017). However, the scaling relies on international averages across all participating countries and so categories may not adequately reflect differences within the South African context specifically. For example, using the international thresholds to describe 'school composition by student background', almost all South African schools fall into the "disadvantaged" category on the derived scale, allowing little differentiation between levels of advantage for schools within South Africa.

Another concern for using responses from questionnaire data is the potential for unreliable self-reported information from students, parents, teachers or school leaders (van Staden et al., 2019). While utilising secondary data like PIRLS enables a detailed exploration of contextual influences on reading achievement in South Africa, concerns about the relevance of questionnaire scaling and potential self-reporting biases need to be kept in mind when interpreting the findings presented in the subsequent papers.

4.4.2 Methodological approaches

The regression analyses and IRT-based analyses used in this study all depend on having high quality data that meets certain assumptions around sample representation, appropriate dimensionality, model-data fit, and reliability of score estimates. However, there are potential limitations with the PIRLS data for the South African sample that need to be considered.

Firstly, the smaller sample size for the Sepedi language group compared to the English and Afrikaans groups introduces uncertainty around the reliability and stability of estimates for the Sepedi group across analyses. This limitation is compounded by the homogeneity observed within the Sepedi group, where students had consistently low performance in PIRLS. This lack of variability makes it challenging to accurately estimate parameters, detect subtle differences within the group, and make robust comparisons with other language groups. With a larger sample or more comparable samples across groups, parameter estimates could potentially differ. Additionally, since the Sepedi group had the lowest overall achievement results in PIRLS, challenges around adequate targeting of item difficulties to student abilities may compound analytic challenges for this group. However, while these factors present analytical challenges, the inclusion of the Sepedi group was crucial for this study as it provided valuable insights into the methodological and practical issues that might arise when dealing with large performance gaps in ILSA data, particularly those related to low performance.

Secondly, while the IRT models selected do align with PIRLS' approaches, there are some differences taken in this research, most notably that sample size limitations prevented the use of the 3PL model for multiple-choice items. However, by using the 2PL model with fewer parameters there is increased stability of estimates which is important given the available sample size.

Finally, the limited variability in overall scores for this dataset may reduce the power and accuracy of DIF methods. DIF analysis methods ideally rely on having adequate score variation across the ability spectrum to reliably detect differences in item functioning between groups (Ercikan, 1998). The low overall achievement levels of South Africa, therefore, raise potential concerns about limited power and accuracy of the DIF approach. Although the limited variability is not optimal for DIF detection, by applying more comprehensive DRF approaches that consider effect sizes and overall impact of DIF, there is still potential for a meaningful evaluation of fairness issues to consider for PIRLS in South Africa.

The concerns and limitations discussed here, particularly regarding the threats to reliability and validity introduced by the low overall achievement levels in South Africa, are a central motivation for the investigations undertaken in this thesis. While these factors may potentially undermine some of the analytical approaches taken, they also underscore the crucial need for careful examination of the PIRLS results through this critical and contextualised lens. By explicitly grappling with these challenges, this research endeavours to provide a more comprehensive and rigorous evaluation of PIRLS within the South African context. In doing so, this research aims to contribute to approaches for the evaluation of large-scale educational assessments, particularly in contexts where overall student performance is very low.

CHAPTER 5 (PAPER 1)
LANGUAGE MATTERS: PIRLS READING ACHIEVEMENT IN SOUTH AFRICA

Abstract

This paper investigates the extent to which factors related to both language and context contribute to students' reading achievement in PIRLS 2016 in South Africa. Multiple linear regression analysis was conducted using data from 12,810 Grade 4 students from 11 language groups to examine relationships between reading achievement and student, home, school, and language factors. The findings indicate that overall PIRLS reading scores were predicted by contextual factors related to students' home resources, students' attitudes towards reading, and school factors such as school location, with 41% variance explained. When test language factors were added, the context and language factors collectively explained 54% of the variation in overall reading scores. The language in which students wrote PIRLS accounted for the most substantial influence on their reading achievement scores. Students taking the assessment in African languages showed substantially lower scores than English or Afrikaans groups, even after adjusting for contextual factors. The findings underscore the limitations of considering just surface-level comparisons of reading achievement and echo calls for context-specific research in developing countries.

Key words

PIRLS, reading achievement, contextual influences, cross-language reading assessment

5.1 Introduction

Reading is a fundamental foundational skill, not just for improving educational outcomes, but for integration into, and contribution to, broader society. As such, supporting and monitoring the development of reading skills have become global priorities (United Nations, 2015). The Progress in International Reading Literacy Study (PIRLS) is an international large-scale assessment (ILSA) of reading comprehension designed and administered by the International Association for the Evaluation of Educational Achievement (IEA). ILSAs like PIRLS aim to evaluate education systems and provide information about factors impacting students' reading achievement across multiple countries. To do this in a way that accounts for the conceptual and contextual complexity, ILSAs develop comprehensive assessment frameworks to guide instrument design that aim to be sensitive to the multifaceted nature of educational achievement as well as the diversity of global educational settings.

The PIRLS assessment framework is founded on a concept of reading as an interactive, constructive process of making meaning from text, integrating skills in decoding, comprehension, and critical evaluation of texts (Mullis et al., 2015a). PIRLS assesses students’ reading skills through an achievement test that consists of different reading passages and items that target various aspects of the reading process across a range of difficulty levels. PIRLS also includes background questionnaires to students, parents, teachers, and school administrators to gain insight into the broader context surrounding reading development. Together, the achievement and questionnaire data aim to provide comprehensive information to inform educational policy and practice (Martin, Mullis, & Foy, 2017). In addition to gathering achievement and background data, PIRLS also conducts detailed analyses of the results to provide insight into ‘what makes a good reader’ across all participating countries (Mullis et al., 2017b).

Research suggests that reading proficiency is not only a matter of individual ability or skills, but is also influenced by the social, economic, cultural and linguistic contexts with which a reader interacts (Castles et al., 2018; Hulme et al., 2022). These sociocultural contexts shape the development of the multiple skills that underpin reading development including linguistic knowledge, background knowledge, decoding skills, metacognitive strategies, and motivation to read (Johnson & Perry, 2012). Empirical studies consistently demonstrate that there are strong associations between reading achievement and factors such as language use at home, school environment and resources, pedagogical techniques, socioeconomic status, and cultural and linguistic practices (e.g., Burgess et al., 2002; Guzmán-Simón et al., 2020; Netten et al., 2014, 2016; Park, 2008; Perez et al., 2004).

However, although often considering the impact of inequality and socioeconomic disparities on reading achievement, a majority of these studies focus primarily on high performing, developed countries (e.g., Araújo & Costa, 2015; Cordero & Mateos-Romero, 2021; Costa & Araújo, 2018; Heppt et al., 2022; Krijnen et al., 2021; Luo et al., 2020; Wang & Wu, 2023). Meanwhile, noticeably less attention is given to developing contexts where gaps in income and access to resources are substantially wider (Nag et al., 2023; Wagner, 2017). Furthermore, there is evidence that shows that although the key factors impacting reading achievement are reasonably consistent across high-performing, high-income countries, they can vary considerably when looking within countries with low achievement (Lagravinese et al., 2020; Marôco, 2021; Park, 2008; van Staden et al., 2019). Typically, low performing countries are less economically developed, with wider gaps in economic inequality, as well as diverse, dynamic, and complex contextual variations across groups within the country (UNESCO, 2014). This means that relying on generalised findings from all countries may lead to incorrect inferences being drawn that are not reflective of the situation within certain countries.

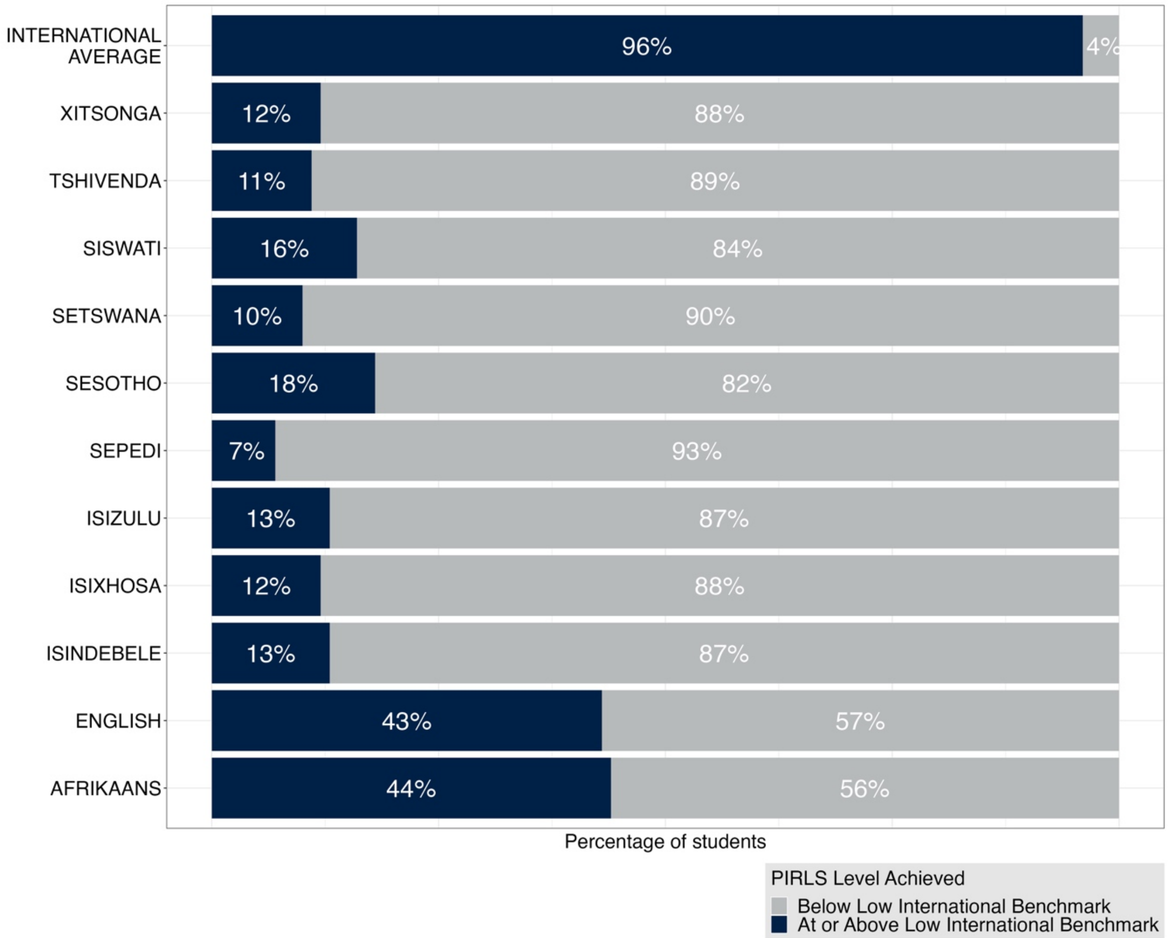
One of the countries with the widest achievement gaps in PIRLS results is South Africa, a lower-middle income country that continues to grapple with substantial socioeconomic inequality that impacts equitable access to education (Leibbrandt & Pabón, 2021; Spaull & Jansen, 2019; United Nations, 2020). As a diverse nation embracing multiple languages and cultural groups, South Africa faces additional complexities in addressing the wide gaps in learning outcomes across groups within the country (Gustafsson & Taylor, 2022; Motala & Carel, 2019; van der Berg & Gustafsson, 2019).

5.1.1 PIRLS in South Africa

South Africa has participated in PIRLS since 2006 and has been the lowest performing country in every cycle out of more than 50 participating countries. In 2016, South Africa had an average overall reading score of 320, well below the international centerpoint of 500. Added to this, in each cycle the range of achievement within South Africa has been wider than the range of achievement in any other participating country. For instance, a high performer like Hong Kong had an average score of 572 with scores ranging from 278 to 761 points. Whereas South Africa, the lowest performer, had scores ranging from 5 to 710 points.

The consistently poor achievement outcomes of South Africa in PIRLS have led to discussions about the depths of the “reading crisis” in the country and intensified efforts to understand the factors influencing reading achievement (Spaull, 2022; van Staden & Combrinck, 2023). However, interpreting the meaning of large-scale assessment results can be complex, especially in highly unequal contexts like South Africa where identity, language, and educational opportunity still align with historical policies of oppression and segregation. This presents challenges because the diversity, inequality, and wide attainment gaps between groups in South Africa mean that average scores can mask substantial variation across subgroups. Indeed, the complex heterogeneity of the results means that looking only at average achievement may result in unfair and often misleading inferences (Spaull & Jansen, 2019).

To assist with the interpretation of the scores, PIRLS sets out four benchmarks of reading achievement: low, intermediate, high, and advanced at set points along its overall reading scale. Students who do not reach the low benchmark are not able to retrieve basic information from the text and consequently are deemed not able to read for meaning. In South Africa, 78% scored below the low benchmark level, this means that only 22% of students were able to demonstrate the ability to read at a basic level after four years of formal schooling (Howie, Combrinck, Roux, Tshele, et al., 2017). The sombre picture of the reading abilities of students in South Africa is even more stark when considering the differences in achievement across the 11 official language groups. Figure 5-1 shows the proportion of students who reached the lowest benchmark internationally, as well as the proportion in each language group in South Africa.



Source: IEA International Data Repository for PIRLS 2016
Note: Sepedi[6]

Figure 5-1 PIRLS 2016 international benchmark achievement by language group in South Africa

The discrepancy between the achievement of students in English and Afrikaans groups when compared to the nine African language groups is stark. In the English and Afrikaans groups 43-44% of students were able to read for meaning, whereas in the nine indigenous African language groups no more than 18% of students were able to demonstrate that they could read for meaning. The challenges with drawing inferences and comparisons on the basis of language in South Africa are numerous and overall results such as these are not sufficient to cover the intricacies of the situation. One of the primary issues often raised is whether or not the test language is the same language

[6] It is important to note that within Sepedi speaking communities some feel that Sepedi was erroneously declared as an official language as it is a dialect of Sesotho sa Leboa (Northern Sotho). (Constitutional Review Committee, 2017)

students speak at home (Akkakoson, 2013; Howard et al., 2014; Yamashita, 2002). In almost every group a majority of students (73%-80%) stated that they always spoke the test language at home. Only the English group, which is the highest achieving group, had a much lower percentage of students (24%) who always spoke the test language at home.

Students in South Africa took the PIRLS reading test in the language that was their medium of instruction for the first three years of primary school. In South Africa, language use at school is guided by five main language policies and regulations (DBE, 2011, 2013; South African Schools Act No. 84 of 1996; Language in Education Policy, 1997; Constitution of the Republic of South Africa, 1996). The dynamics of balancing language rights, healing past injustice and providing equitable opportunities with more global notions of 'best-practice' for multilingual education is something that South Africa's current language in education policies are attempting to address. They provide a set of guidelines regarding how languages are chosen as media of instruction and as subjects at each school, as well as requirements for the number of languages taught, which languages can be selected, and the time allocation given to each language in the classroom.

Any of the 11 official languages may be selected as a medium of instruction as well as South African Sign Language, which has recently received approval to become the 12th official language (Parliament of the Republic of South Africa, 2023). In an effort to emphasise the importance of building a solid foundation in the mother-tongue, the policy encourages a 'home language' to be used in the first three years (while learning English as a language subject) and then proposes that schools should switch to English as the medium of instruction from Grade 4. Most schools follow this recommendation and start teaching in the most spoken language in the province for the first three years of primary school and then switch the medium of instruction to English in Grade 4, or to Afrikaans in a few schools. These policies are controversial and selecting languages can be a difficult and complicated endeavour (Broom, 2004; Gustafsson, 2018; Letseka, 2014; Spaull et al., 2020; Zimmerman, 2014). Regardless of one's position in the debates, however, it remains clear that challenges surrounding language in education continue to impact student outcomes in South Africa.

South Africa's context of widespread poverty, extreme inequality, and multiple official languages presents a particularly complex environment for reading development. Moreover, the legacy of apartheid policies continues to disproportionately impact access to resources and quality education for the majority of students. Given these intricacies, considering the average results across groups may not accurately capture the nuances of the situation, particularly as there are potentially different factors at play influencing reading achievement in South Africa compared to more developed countries. Having within-country insights about the factors influencing reading development in South Africa is an important step in gaining a nuanced picture of the situation. With a more detailed understanding of the factors at play, targeted interventions can seek to address the concerningly low reading achievement by targeting relevant, context-specific issues.

5.1.2 Aims and research questions

This paper aims to understand how different contextual and linguistic factors influence students' PIRLS 2016 reading achievement in South Africa. It focuses on understanding disparities in performance across language groups and disentangling the complex relationships between language, socioeconomic and contextual factors, and achievement. Three research questions guide the investigation:

1. To what extent do contextual factors impact students' overall PIRLS 2016 reading achievement in South Africa?
2. To what extent do factors related to the use of the test language at home impact students' overall PIRLS 2016 reading achievement in South Africa?
3. To what extent does the test language impact students' overall PIRLS 2016 reading achievement in South Africa?

5.2 Literature review

5.2.1 Reading comprehension

Research that endeavours to understand what reading is and what it involves has been widely conducted across multiple disciplines. There have been numerous approaches, conceptualisations and definitions of reading over the years. As PIRLS is a test of reading comprehension, this review focuses on a discussion of conceptualisations and contributions specific to reading comprehension. Central to these is an understanding that reading comprehension is more than a matter of decoding symbols into sounds, it involves building an understanding of meaning from texts. The PIRLS 2016 reading framework sees reading as "the ability to understand and use those written language forms required by society and/or valued by the individual" (Mullis et al., 2015b, p. 12). The framework centres active meaning-making as fundamental to comprehending text and describes the process of constructing meaning from written language as an interactive process that coordinates multiple skills and capacities. This aligns with decades of research that has suggested that skilled reading comprehension integrates diverse linguistic, cognitive, motivational and sociocultural factors rather than a single skill (Dooley, 2017; Duke & Carlisle, 2011; Huey, 1908; Kintsch & van Dijk, 1978; Laufer, 1989; Thorndike, 1917).

One influential reading framework is the simple view of reading, which emphasises that reading comprehension depends specifically on language comprehension abilities and decoding skill (Gough & Tunmer, 1986; Hoover & Gough, 1990). The influence of context on constructing meaning in reading comprehension is also strongly regarded, with theorists noting that contextual influence can be both beneficial and detrimental to reading comprehension (Christianson & Luke, 2011; Cummins, 2012; Miller & Faircloth, 2009; Rueda, 2013).

Following on from these established perspectives, recent research on reading comprehension reflects a growing focus on developing a more in-depth understanding of multilingual reading development. Much of the early research into reading development centred around learning to read in English, either as a first or, less frequently, as a second language. While the role of the first language on reading development in additional languages has featured in reading research for a number of decades (Alderson, 1985; Bernhardt, 1993; Vermeer, 2004); more recent studies seek to understand the development of reading in the more realistic, dynamic, and complex multilingual contexts that exist today (Coady et al., 2022; Kretzer & Kaschula, 2021; Makalela, 2017; McBride, 2017; Miller et al., 2018; Murphy, 2018; Pae & Sevcik, 2017). As perspectives on reading comprehension expand to encompass more diverse linguistic contexts, contemporary research aims to build integrated models that capture the complex interplay between the skills, cognitive processes, and contextual factors that influence reading comprehension (Castles et al., 2018; Goldenberg, 2020; Ouellette & Beers, 2010; Snowling et al., 2019).

5.2.2 Assessing reading comprehension

The multiple factors associated with reading comprehension make developing assessments to measure it effectively and holistically a challenging undertaking. Moreover, isolating reading skills from language and conceptual knowledge is an extremely difficult task, particularly because language is both part of the construct being measured and the medium through which the assessment takes place. To this end, test developers create comprehensive assessment frameworks that build on theoretical models of reading comprehension. These frameworks describe the relevant components, constructs, processes, and difficulty levels targeted in their reading assessment as well as how they incorporate considerations of the influence of broader contextual factors (e.g., Mullis et al., 2015a; OECD, 2023).

An additional consideration of ILSAs is that the assessment is translated to different languages. PIRLS 2016 for example, was conducted in 50 countries and translated into over 40 different languages (Mullis et al., 2017a). The procedures involved in test adaptation for ILSAs are thus carefully designed and regulated in an attempt to consider the educational, linguistic, cultural, and psychometric properties that may affect different versions of a test (Bartram & Hambleton, 2016; De Boeck & Elosua, 2016; Rutkowski et al., 2013).

PIRLS assesses two aspects of reading – the purposes for reading (literary and informational) and the comprehension processes involved in reading (i.e., retrieve explicit information, make straightforward inferences, interpret and integrate ideas and information, examine and evaluate content, language, and textual elements). The reading test consists of 12 reading passages and 183 associated items, with students each seeing two passages and answering around 30 items (~15 per passage). Passages range from 400-800 words and cover a variety of topics and text types selected to reflect authentic and age-appropriate reading materials. As the impact of background knowledge on reading comprehension is recognised by the PIRLS reading framework, the passages go through a thorough selection, translation and adaptation process to ensure the texts are relevant to students internationally (Martin, Mullis, & Hooper, 2017).

The interaction and influence of the broader social context on reading comprehension is another core feature of the PIRLS Literacy 2016 reading framework (Mullis et al., 2015b). To account for this influence, five background questionnaires (home, student, teacher, school, and curriculum) are administered alongside the reading assessment (Hooper et al., 2015). The questionnaires cover a range of topics considered to be associated with reading development such as home support for learning, curricula, teacher education, and classroom practices and are adaptable according to the needs of each country. In South Africa, for example, the questionnaires were adapted to include questions related to the location of schools, and the relationship between the test language and the language(s) spoken at home (Foy, 2018b). The subscale structure of the reading test, together with the detailed contextual questionnaire information allows for the analysis of factors predicting different comprehension skills.

5.2.3 Factors influencing reading comprehension

Research and empirical investigations into reading have identified a several factors that influence the successful development of reading comprehension skills. These factors can be broadly categorised into individual student characteristics, features of the home environment, as well as aspects related to teaching and learning at school (Hulme et al., 2022; Melby-Lervåg & Lervåg, 2014; Westerveld et al., 2020).

Student characteristics

Substantial research identifies that student characteristics related to cognitive ability, linguistic skills and motivational factors all support reading achievement (e.g., Evans, 2011; Miller & Faircloth, 2009; Murphy, 2018; Taboada et al., 2009; Wigfield et al., 2016). One of the fundamental skills students need to have in order to become successful readers is oral language proficiency (Lervåg et al., 2018). Key linguistic skills for early reading development include decoding, phonemic awareness, morphological awareness, and vocabulary knowledge (e.g., Baumann, 2014; Clarke et al., 2010; Murphy, 2018; Wilsenach, 2015). The decoding skill enables efficient word recognition which in turn provides essential access to text meaning (Melby-Lervåg et al., 2012; Muter et al., 2004). Another essential skill which is integral to becoming a successful reader is vocabulary size; studies across multiple contexts have demonstrated that vocabulary size strongly predicts reading achievement for both first and second language learners (Kan & Murphy, 2020; Sénéchal et al., 2006; Soares De Sousa, 2011; Stoffelsma, 2019; Wilsenach, 2015). The more words students know, the more likely they are able to effectively, and efficiently, understand the text.

The role of a student's language proficiency in developing the foundation for reading comprehension cannot be understated (Hulme & Snowling, 2013; Lervåg et al., 2018). However, attitudes towards reading, motivations to read, as well as confidence and engagement in reading tasks also show positive relationships with reading achievement for students in some contexts (Miller & Faircloth, 2009; Rueda, 2013; Toste et al., 2020; Wigfield et al., 2016).

Home factors

The home literacy environment, including access to books and parental reading behaviour, provides important early opportunities to develop language skills and positive reading behaviour (Burgess et al., 2002; Guzmán-Simón et al., 2020; Nag et al., 2023; Sénéchal & Lefevre, 2014).

However, children from lower-income households are likely to have less access to reading materials, and their parents are less likely to have strong literacy skills themselves (Buckingham et al., 2014). This creates a socioeconomic achievement gap, as higher income homes can provide more resources and support for developing reading skills (Nag et al., 2023).

The language environment at home is another critical feature influencing reading development. When the language(s) used at home differ from the school instructional language, oral skills in the home language may not directly transfer to reading in another language (Dixon et al., 2012; Goodrich et al., 2013; Probert & De Vos, 2016; Wilsenach, 2013). Differences in status between home and school languages can also negatively influence student attitudes and motivation to read (Cele, 2001; Hoff, 2013; Howard et al., 2014; Nag et al., 2023). Proactive language policies and instructional approaches that leverage multilingualism may help promote reading success (Taylor et al., 2008). However, the complex sociolinguistic dynamics between home and school language environments can profoundly impact the trajectories of reading development for students.

School and teacher factors

Research shows that effective, skills-based reading instruction and curriculum positively impact reading development (Castles et al., 2018; Compton-Lilly et al., 2023; McKeown et al., 2009; Pearson, 2004, 2011; Reutzel et al., 2014). Teachers can also play a key role – factors related to expertise, preparation, and specialised instruction training have been shown to improve reading achievement (Blair et al., 2007; Didion et al., 2020; Podhajski et al., 2009). However, in low- and middle-income countries research has shown that factors like socioeconomic status, home environments and school resources may outweigh individual teacher factors (Aturupane et al., 2013; Barrett & Tikly, 2013; Mohangi et al., 2016; Nel et al., 2016). It is important to note that the limited empirical evidence for low- and middle-income countries showing the impact of teacher quality relative to other socioeconomic factors does not negate the importance of high-quality instruction for reading development. Rather, the reduced empirical evidence for teacher or school quality effects in low- and middle-income countries may be partly attributed to less variance in teacher quality at a within-country level, as teacher preparation and support systems are likely to be similarly constrained for all teachers. Additional difficulties in defining and measuring teacher quality across different cultural and educational contexts are also likely to contribute to this complexity. Hanushek and Rivkin (2006) argue that conventional measures of teacher quality such as certification or years of experience, often explain little of the variation in teacher quality as reflected in student achievement, highlighting the need for more nuanced measures of teacher effectiveness in diverse educational contexts.

Despite the challenges determining what constitutes high-quality teaching, targeted interventions to improve teacher pedagogy and practice in developing contexts have shown promising results. Studies in several developing countries (including South Africa) have demonstrated that when teachers receive support in implementing effective instructional strategies, particularly those tailored to local contexts, student achievement can improve significantly (Piper et al., 2018; Kotzé et al., 2019). Furthermore, large disparities in access to quality instruction widen achievement gaps; this is particularly true in countries where access to quality instruction varies substantially across groups (Lagravinese et al., 2020; Park, 2008).

Interactions between factors

While student characteristics, home environment, and school or teacher factors are described here as distinct categories, it is essential to recognise that these elements dynamically interact and shape one another in important ways. For example, higher family income enables greater access to reading materials at home as well as higher quality schools, which collectively provide more support for developing strong reading skills (Nag et al., 2023; Park, 2008; van Bergen et al., 2017). Conversely, limitations in the home literacy environment coupled with inadequate reading instruction at under-resourced schools can compound challenges for students from disadvantaged backgrounds (Chansa-Kabali et al., 2014; Hoff, 2013). Furthermore, student attitudes and motivation are influenced by home reading activities, teacher practices as well as the complex dynamics of language and social status (Araújo & Costa, 2015; Howard et al., 2014; Sénéchal &

Lefevre, 2014). This means that building a comprehensive understanding of what leads to successful reading comprehension involves considering not just the factors involved, but also the interdependencies between these factors.

5.2.4 Limitations of current evidence base

A substantial limitation in the existing literature on reading comprehension is that a number of the methods used to collect data disproportionately focus on developed, high-income countries (Howie, 2012, 2022; Luo et al., 2020; Murphy & Unthiah, 2015; Nag et al., 2023; Wagner, 2017). Far less research examines reading development in disadvantaged or developing countries, leading to an unbalanced view of reading comprehension across different contexts. Added to this, the vast majority of countries participating in ILSAs are high performing, developed countries, which leads to further issues when considering the results from these tests (Addey & Sellar, 2019; Atar et al., 2023; Rutkowski & Rutkowski, 2021, 2022; Wagemaker, 2020). The overrepresentation of high-income, high-performing countries in the participant pool can impact the validity of information that can be drawn from the results, as well as the generalisability of the findings across contexts (Atar et al., 2023; Rutkowski et al., 2019; Tijmstra et al., 2020).

Over 40 years ago, Heyneman and Loxley (1982) already cautioned that influences on academic achievement differ markedly between high-income and low-income countries. They found that variables highly predictive of achievement in low-income countries may be overlooked in analyses of international assessments which have disproportionately more high-income countries in their samples. Consequently, findings of "no impact" on achievement overall frequently mask strong effects present in low-income countries. For example, although IEA data from 1960-1980 suggested that school quality does not have a strong influence on achievement globally, Heyneman and Loxley showed this relationship actually is significant in low-income countries (Heyneman & Loxley, 1983).

Recent evidence confirms that there are substantial disparities between the factors that influence academic outcomes in low and high achieving countries (Ahmed et al., 2022; Gomendio & Wert, 2023; Luo et al., 2020; Marôco, 2021; Nag et al., 2023; van Staden & Combrinck, 2023; Wagner et al., 2011; Wagner & Castillo, 2014). While some influences remain consistent globally, transferring findings directly from highly developed, homogenous settings risks overlooking needs or features unique to diverse and disadvantaged populations. More localised studies also reveal variation in key factors at a within-country level when comparing performance across groups within specific contexts (Chiu & Chow, 2010; Govender & Hugo, 2020; Gustafsson, 2019b; Naidoo et al., 2014; Nel et al., 2016; Pretorius & Spaull, 2016; Spaull & Pretorius, 2019; van Staden et al., 2019). Also important to note is that when studies do focus on developing countries, samples frequently include only advantaged, urban schools and students (Andringa & Godfroid, 2020; Scholtz, 2021). Yet the severe inequality in developing countries means rural schools and poorer households are likely to have distinct challenges affecting achievement and without representative data, assumptions based on subgroups will further distort understanding. This is one of the benefits of considering the PIRLS results in South Africa; PIRLS is one of few assessments in the country that collects nationally representative data stratified by language of instruction, school type and income level (Howie, Combrinck, Tshele, et al., 2017).

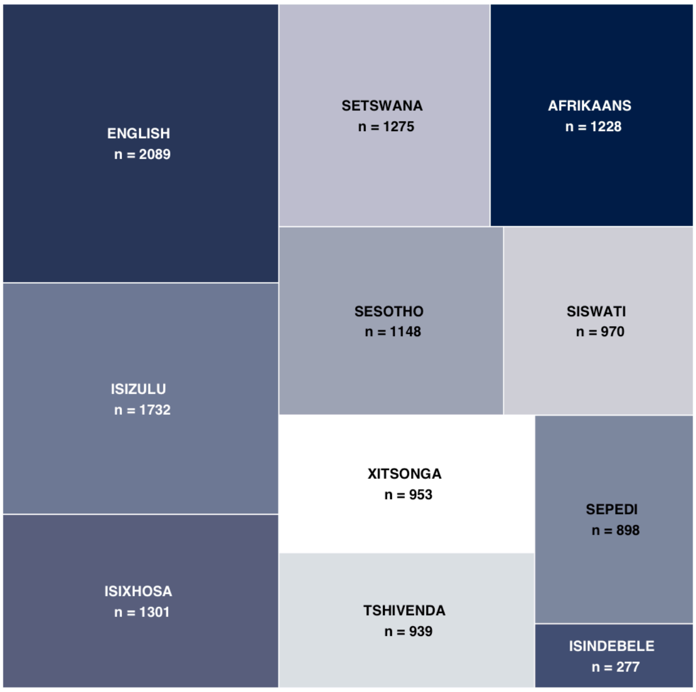
Ultimately, the considerable economic and social differences between most developed countries with an established evidence base and under-studied developing contexts signal a need for dedicated, context-specific research. Rather than generalising from findings across all contexts, researchers must carefully examine results within lower performing countries to identify factors that hold significant influence on academic achievement in those settings, without inaccurately concluding they have no impact overall. This study helps address this gap through a context-focused analysis of reading achievement patterns and associated factors within South Africa specifically.

5.3 Methods

5.3.1 Data

This paper features data from Grade 4 students in South Africa who participated in PIRLS Literacy 2016 and includes information about the students’ achievement results for PIRLS and their background questionnaire responses. The data also includes questionnaire responses from the home, teacher and school questionnaires responded to by students' primary caregivers at home, their teachers, and principals at the schools they attended.

The PIRLS 2016 national dataset for South Africa (N=12,810) is a nationally representative sample of students in Grade 4 from all 9 provinces, and 11 language groups and was retrieved from the publicly available IEA international data repository. Figure 5-2 shows a treemap of the dataset with blocks portraying the proportions of students from each language group. The English group had the most students (2,089), followed by isiZulu (1,737), isiXhosa (1,301), Setswana (1,275), Afrikaans (1,228), Sesotho (1,148), siSwati (970), Xitsonga (953), Tshivenda (939), Sepedi (898) and isiNdebele (277).



Source: IEA International Data Repository for PIRLS 2016

Figure 5-2 Distribution of students per language group in the PIRLS 2016 South African dataset

PIRLS 2016 used a two-stage random sampling design to select schools for participation in the international study (Martin, Mullis, & Foy, 2017). In the first stage, a sample of schools is selected in each country and in the second stage, one or two intact classes from a specific grade-level are selected from each of those schools. PIRLS 2016 selected whole classes rather than individual students because of their focus on the curricular and instructional experiences related to reading achievement and to ensure minimal disruption to the day-to-day school routine (Martin, Mullis, & Hooper, 2017). The South African sample in this study is nationally representative, stratified by language and province (Howie et al., 2017). However, the implicit stratification in the two-stage sampling design of PIRLS means that there is a higher number of English classes (and therefore students) compared to the other language groups. This is due to the complex multilingual nature of schools in South Africa, where a single school may offer multiple language streams. It is typical for one of these options to be English, while the other language option(s) vary greatly by location. This means that there is a higher number of English classes, and by extension students in the English group than in other groups in the sample.

5.3.2 Multiple linear regression

This paper explores the role that contextual and linguistic factors play on overall PIRLS Literacy 2016 performance in South Africa with a specific focus on the relationship between test language and reading achievement. The comprehensive data from the PIRLS background questionnaires enables the exploration of a number of variables related to students' language backgrounds. Across the student and home questionnaires there are questions which delve into the home language situation of the student by asking both students and parents about the frequency at which the test language is spoken at home, as well as what language each parent uses to communicate with the student at home. The home and student questionnaires also cover other contextual issues that have been shown to be relevant in reading development such as indicators of socioeconomic status, attitudes towards reading, and early literacy experiences.

However, although the integral role of language and context on reading development is well recognised (Landi, 2012; McClung & Pearson, 2019; Perfetti & Harris, 2013; Pretorius & Mampuru, 2007; van Staden & Combrinck, 2023); exploring this situation can be challenging due to the complex interplay between influential factors. For example, when examining contextual factors in isolation it may be seen that there is a strong relationship between school type (e.g., rural or urban) and reading achievement, however, when background variables are examined concurrently, it may become evident that school type is highly associated with the socioeconomic status of students. In this scenario, school type may be functioning as a proxy for socioeconomic status and the results could therefore be misleading. Therefore, to avoid an inaccurate estimation of the effect of contextual and language variables on reading performance, multiple linear regression analysis was used as it is able to account for interactions between potentially confounded variables across groups (Grömping, 2012). By using multiple linear regression to examine the relationships; several variables can be explored simultaneously while accounting for the interactions between them. This approach has also been applied by similar studies considering the influence of variables from background questionnaires on PIRLS results (e.g., Guzmán-Simón et al., 2020; Taylor & Yu, 2009; van Staden et al., 2019).

It is important to note that although multilevel modelling, which accounts for the nested structure of a dataset, might seem appropriate for this analysis, PIRLS selects one whole class at each school leading to the conflation of school and class levels. Furthermore, to account for the sampling design of PIRLS, sampling weights for student, school and teacher variables are provided for secondary investigations, and accounted for in the computation of achievement and background data in PIRLS (Martin, Mullis, & Hooper, 2017).

Variables

The dependent variable in this analysis is the overall achievement of each student in PIRLS Literacy 2016. Overall achievement for PIRLS is represented by a set of five plausible values (PVs) rather than one average score. PVs are randomly drawn values from the posterior distribution of potential scores for each student based on their responses to the assessment items. PVs are commonly used in ILSAs to represent overall scores as they account for the measurement error inherent in estimating student achievement based on a sample of assessment items, and therefore present a more accurate estimation of overall achievement (Wang, 2020).

The predictor variables in this analysis are from the PIRLS 2016 contextual questionnaire data for South Africa. The variables from the questionnaire data represent responses to individual items as well as scaled items provided by PIRLS for sets of items that measure one underlying construct (for a description of the item sets related to each scale used in this analysis, see Appendix C-1). A set of 21 predictor variables were selected, separated into 15 contextual variables and 6 language variables. The contextual variables represent relevant student, home, and school variables; while the language variables relate to language use at home as well as test language. Variables were selected for inclusion as predictors if they were shown to have an impact on reading achievement in PIRLS internationally in the international results for PIRLS 2016 presented by Mullis et al. (2017). Additionally, variables were included if they were found to be contextually relevant factors in South Africa presented by Howie et al. (2017).

Although the results from both international and local studies show that teacher factors such as job satisfaction and specialised training may impact the educational outcomes of students; teacher variables were not included in this model due to validity and reliability concerns with the data from South African teacher questionnaires in PIRLS 2016 raised by van Staden et al. (2019). Their findings revealed overly positive reporting by teachers and as such they cautioned against reliance on data from the teacher questionnaires, particularly in developing contexts. Additionally, findings from similar studies in low- and middle-income countries demonstrate that socioeconomic disparities both at home and school can outweigh the influence of teacher characteristics (Aturupane et al., 2013; Mohangi et al., 2016). It is important to note, however, that the exclusion of these variables does not imply that teacher quality is unrelated to student outcomes, but rather that there are challenges in accurately measuring and representing these factors in the current dataset.

A short description of each of the variables selected as predictors for this study, as well as the contextual questionnaire in which they are sourced is shown in Table 5-1.

.

Table 5-1

Description of predictor variables and their source questionnaires

Category	Code	Variable	Questionnaire
Contextual			
Control	itsex	Sex of students	-
Student	asdgssb*	Students sense of school belonging	Student
	asdgsb*	Student experience of bullying	Student
	asdgerl*	Students engaged in reading lessons	Student
	asdgslr*	Students like reading	Student
	asdgscr*	Students confident in reading	Student
Home	asbghrl*	Home resources for learning	Home & Student
	asdhela*	Early literacy activities at home	Home
	asdhelt*	Could do early literacy tasks before school	Home
	asdhplr*	Parents like reading	Home
School	acbg05b	School location	School
	acdgrrs*	Instruction affected by reading resource shortages	School
	acbgeas*	School emphasis on academic success	School
	acdgdas*	School discipline	School
	acdg03	School composition by student background	School
Language			
Language use	asbg03	Frequency of test language spoken at home	Student
	asbh04a	Child spoke test language before beginning school	Home
	asbh17	Frequency child speaks test language at home	Home
	asbh21aa	Father speaks test language to communicate with child at home	Home
	asbh21ab	Mother speaks test language to communicate with child at home	Home
Test language	itlang	Test language	-

*Note: * derived scaled items*

5.4 Analysis

A forward approach to multiple linear regression modelling was used to understand the relative impact of the different predictor variables. This approach was chosen because it aligns with the conceptual framework guiding the analysis, which grouped predictors into distinct categories according to their potential influence (home factors, student factors, etc.). Additionally, a hierarchical forward modelling strategy enabled an exploration of the specific contributions of each variable set, while guarding against the inclusion of variables that may diminish model fit (Cohen et al., 2013). Predictor variables were categorised according to the questionnaire in which they occurred; i.e., home factors, student factors, school factors, as well as an additional category of ‘language factors’ which includes variables specifically related to language use from across the questionnaires.

First, a control model (Model 0) was estimated that included gender as a control variable in order to account for well-established differential gender effects on reading achievement related to socioeconomic status (e.g., Baye & Monseur, 2016; Stoet & Geary, 2013). Following that five multiple linear regression models were estimated adding variable sets sequentially. Model 1 included four variables related to the students’ home environment. Model 2 added five variables that described students’ attitudes towards learning and school. Model 3 added five variables from the school questionnaire that related to the school climate, as well as school location and socioeconomic indicators of the school. The final two models added variables related to language factors. Model 4 added five variables that captured the extent to which students use the test language at home, and Model 5 added the test language variable. Building the regression models in this way enables an exploration of the specific contributions of home language factors (Model 4) and test language (Model 5), over and above contextual factors alone (Models 1 – 3). Details of variables included in each model are presented in Table 5-2.

The regression analyses were conducted using the EdSurvey package in R as it is designed specifically for analysis of large-scale assessment data and is capable of accounting for the complex sample designs, replicate weights, and PVs in the analysis of PIRLS data (Bailey et al., 2023). As recommended for within country analyses of PIRLS data such as this, student weights were applied so that subgroups were represented proportionally (Rutkowski et al., 2010). Variance introduced by sampling was addressed through use of the Jackknife repeated replication (JRR) technique with 150 replications, as recommended for analysis of PIRLS achievement data (Foy, 2018a).

The relationships between the set of predictor variables and PIRLS reading achievement were examined in each regression model using three key statistics (R^2, B coefficients, β coefficients). The magnitude of the collective relationships was determined using multiple R^2 values, which reflect the proportion of variance in overall reading achievement scores explained by each model. Predictors demonstrating statistically significant relationships were identified based on t-values greater than 1.96 ($p < .05$). Next, the magnitude and direction of the relationships between individual predictors and achievement were inspected using both unstandardised (B) and standardised regression coefficients (β) (Bailey & Cohen, 2019). The unstandardised coefficients represent the average points difference between comparisons being made, while the standardised coefficients represent the standard deviation change in reading scores associated with a 1-SD increase in predictor variables (Ziglari, 2017). By considering both statistics, the results illustrate the practical magnitude of effects in terms of actual point differences on the achievement scale (B) as well as the relative strength of relationships across variables measured (β).

The reliability estimation of multiple linear regression and accurate interpretation of results depends on a number of assumptions about the data and model being met. Therefore, before proceeding with the analysis in this paper, diagnostic checks were conducted on the regression model to ensure the assumptions of linearity, absence of multicollinearity, homoscedasticity, and normality of residuals were met. Examination of partial residual plots, correlations, variance inflation factors, scale-location plots, histograms, and Q-Q plots showed minimal evidence of assumption violations. The linearity assumption held reasonably well for all predictors, correlations between predictors were moderately low, variance inflation factors were within acceptable limits, a test for heteroscedasticity was non-significant, and residuals showed normal distribution. Further details on these diagnostic checks are provided in Appendix C-2.

5.5 Results

5.5.1 Descriptive statistics

The descriptive statistics for all variables used in the regression analysis are shown in Table 5-2. Continuous variables describe the mean and standard deviation of responses for that variable, while categorical variables describe the percentage of responses in each category. For scaled variables categorical options were used where possible to aid interpretability, however, continuous options were selected for four variables due to low proportions of responses in some categories (*home resources for learning*, *students could do early literacy skills before school*, *parents like reading*, *school emphasis on academic success*, and *reading resource shortages*).

Additionally, three of the categorical variables were recoded to create more suitable categories to aid interpretability and to address sparseness within response options. *Language spoken at home* responses (for parent and students) were collapsed from four into three groups indicating the frequency the test language was spoken at home (always/almost always, sometimes, never). *School composition by student background* was recoded into a simplified distinction between disadvantaged and not disadvantaged schools due to very few responses in the 'affluent' category.

Table 5-2

Descriptive statistics for regression analysis variables

Variable	Response options	N-used	Descriptive
Reading achievement			
Plausible value 1		2479	Mean = 353.7, SD = 104.3
Plausible value 2		2479	Mean = 352.2, SD = 105
Plausible value 3		2479	Mean = 351.3, SD = 105.4
Plausible value 4		2479	Mean = 351.5, SD = 104.2
Plausible value 5		2479	Mean = 352, SD = 104.1
Model 0 – Control			

Variable	Response options	N-used	Descriptive
Student gender	(Girl)	1274	51%
	Boy	1205	49%
Model 1 – Home factors			
Home resources for learning		2479	Mean = 8.5, SD = 1.4
Early literacy activities at home	(Often)	801	32%
	Sometimes	1597	64%
	Never or almost never	81	3%
Could do early literacy tasks before school	(Very well)	860	35%
	Moderately well	953	38%
	Not well	666	27%
Parents like reading		2479	Mean = 9.8, SD = 1.7
Model 2 – Student factors			
Students sense of school belonging	(High sense of school belonging)	1495	60%
	Some sense of school belonging	819	33%
	Little sense of school belonging	165	7%
Student experience of bullying	(Almost never)	551	22%
	About monthly	898	36%
	About weekly	1030	42%
Student engaged in reading lessons	(Very engaged)	1695	68%
	Somewhat engaged	663	27%
	Less than engaged	121	5%
Student likes reading	(Very much like reading)	1463	59%
	Somewhat like reading	823	33%
	Do not like reading	193	8%
Student confident in reading	(Very confident)	599	24%
	Somewhat confident	828	33%
	Not confident	1052	42%
Model 3 – School factors			
School emphasis on academic success		2479	Mean = 9.2, SD = 2.1

Variable	Response options	N-used	Descriptive
Reading resource shortages		2479	Mean = 9.1, SD = 1.4
School discipline	(Hardly any problems)	542	22%
	Minor problems	1304	53%
	Moderate to severe problems	633	26%
School location	(Urban-densely populated)	249	10%
	Medium size city or large town	76	3%
	Small town or village	661	27%
	Remote rural	723	29%
	Suburban/township	770	31%
School composition by student background	(Not disadvantaged)	592	24%
	Disadvantaged	1887	76%
Model 4 – Language use factors			
Frequency of test language spoken at home	(Always/almost always)	1877	76%
	Sometimes	480	19%
	Never	122	5%
Frequency child speaks test language at home	(Always/almost always)	2144	86%
	Sometimes	311	13%
	Never	24	1%
Father speaks test language with child at home	(Yes)	1927	78%
	No	552	22%
Mother speaks test language with child at home	(Yes)	2035	82%
	No	444	18%
Child spoke test language before beginning school	(Yes)	2126	86%
	No	353	14%
Model 5 – Test language			
Test language	(English)	504	20%
	Afrikaans	285	11%
	isiNdebele	99	4%

Variable	Response options	N-used	Descriptive
	isiXhosa	211	9%
	isiZulu	299	12%
	Sepedi	203	8%
	Sesotho	234	9%
	Setswana	197	8%
	siSwati	152	6%
	Tshivenda	174	7%
	Xitsonga	121	5%

Note: ( ) Category used as reference[7] for regression modelling

5.5.2 Regression results

The multiple linear regression analyses produced models with progressively increasing explanatory power in predicting reading achievement. Table 5-3 presents the multiple R^2 value for each model which gives an indication of the percentage of variance explained by the variables in each model. The initial gender-only control model accounted for 6% of variance in PIRLS reading scores. The inclusion of home environment factors raised the explanatory capacity to 23%, while adding student attitudes increased this to 35%. School climate variables further improved prediction to 41% of variance explained. The final two models incorporating language elements had the highest explanatory power, with language use factors increasing R^2 to 45% and the addition of the test language measure boosting overall variance explained in reading achievement to 54%.

An overview of results describing the coefficients from each predictor in all models is presented in Table 5-4 on page 119 (additional tables with the detailed results of each model are included in Appendix C-3). Further analysis of the effect of each of the individual predictors within each model set sheds light on the most influential elements as well as the relative effect they have on reading achievement and are discussed in the sections that follow.

[7] To aid in the interpretation of each category in relation to its effect on reading achievement; the reference for each variable was selected as the category expected to have the most positive association with the dependent variable based on theory and previous research.

Table 5-3

Summary of multiple regression model results

Model	Theme	Variables included	N-used[8]	R^2	Intercept	F-test results
Model 0	Gender	1	12807	0.06	346.7	F(1,75) = 290.1, p < .001)
Model 1	Home factors	5	5586	0.231	178	F(7,69) = 42.9, p < .001)
Model 2	Student factors	10	5157	0.355	280.2	F(17,59) = 29.8, p < .001)
Model 3	School factors	15	4151	0.412	451.6	F(26,40) = 24, p < .001)
Model 4	Language use	20	2479	0.449	449.7	F(33,31) = 15.3, p < .001)
Model 5	Test language	21	2479	0.539	489	F(43,21) = 24.2, p < .001)

5.5.3 Modelling contextual factors

Gender

Model 0 incorporated gender as a control variable. The results showed that gender had a significant influence on reading achievement, with girls scoring an average of 52 points higher than boys (B = -52.1 points, β = -.24, p < .001). This initial model accounted for 6% of the variance in reading scores. Overall, the gender effect aligns with previous research showing girls often outperform boys in reading assessments across contexts (e.g., Hermann & Kopasz, 2021; Marks, 2008; van Hek et al., 2016). Establishing this baseline gender difference is important for understanding how additional variables may mediate or moderate achievement gaps across subsequent models, particularly because the gender gap in PIRLS achievement in South Africa is one of the widest internationally (Mullis et al., 2017a).

Home factors

Model 1 estimated the influence of variables related to students' home environment on their reading achievement. This model significantly predicted reading achievement for South African

[8] The full student sample who participated in the reading assessment was N = 12,810. However, the number of students with valid data for the variables included in each model (N-used) varies as not all students completed the supplementary questionnaires or had matching data from parents, school principals etc.

students (R^2 = .23, F(8,68) = 39.5, p < .001). The R^2 value for this model indicates that 23% of the variance in overall reading achievement for these students can be attributed to factors related to their home. The home factors in this model included variables related to home resources for learning as well as the home literacy environment. The home resources for learning scale is a derived variable based on selected socioeconomic and home environment indicators across both the student and parent questionnaires. It incorporates items such as number of books in the home, number of study supports at home and the parents' education and occupation. The factors related to the home literacy environment included parent self-reports of their own enjoyment of reading, as well as how frequently they engaged in early literacy activities with their child at home (e.g., reading books, telling stories, letter games). The relationships between the individual factors included in the model and students reading achievement are shown in Table 5-4.

All factors that related to students' home environment significantly impacted their reading achievement to some extent (p < .05 level). Considering the standardised coefficients (β) which give an indication of the magnitude of the effect of each variable, home resources for learning and the extent to which students could do early literacy tasks before beginning school had the most substantial influence on reading achievement. Home resources for learning had the largest effect in this model (β = .36); students with more home resources for learning scored over 26 points higher on average (B = 26.7 points, p < .001). Students whose parents reported that they could not do early literacy tasks well when they started school, scored an average of 38 points lower than those whose parents reported they could do those tasks very well when they started school (B = -37.8 points, β = -.15, p < .001).

Although the relationship between reading achievement and parents' attitudes towards reading was significant (p < .05), the magnitude of this effect (β = .05) was smaller than for other variables in this model. Similarly, although the frequency at which parents reported engaging in early literacy activities with their child had a significant influence on reading achievement, the magnitude of this influence was lower than other home factors (-.05 < β < -.02, p <.05).

Student factors

Model 2 estimated the influence of variables related to students' attitudes towards reading and their school environment on their reading achievement. This model significantly predicted reading achievement for South African students ($R^2 = .353$, $F(17,59) = 29.8$, $p < .001$). The R^2 value indicated that 35% of the variance in overall reading achievement can be attributed to students' attitudes towards learning and their school environment. Two variables in this model related to students' experiences of the school environment, these included their sense of school belonging and the frequency at which they experienced bullying. An additional three variables reflected students' attitudes towards reading in terms of how engaged they felt in reading lessons, how much they liked reading and how confident they were in reading.

Confidence in reading had the most substantial influence on reading achievement in this model. Students who reported they were somewhat confident academically scored nearly 48 points lower on average than students who were very confident ($B = -47.8$ points, $\beta = -.20$, $p < .001$). Students who were not confident in reading scored over 80 points lower ($B = -83.5$ points, $\beta = -.37$, $p < .001$). The extent to which students felt engaged in reading lessons also had a significant influence on reading achievement with lower engagement scoring lower overall ($B = -44.1$ points, $\beta = -.09$, $p < .001$). The extent to which students like reading did not show a significant relationship with reading achievement in this model.

Although students' sense of school belonging did not show a significant relationship; the frequency at which they reported experiencing bullying at school did influence their reading achievement. Students who experienced bullying about weekly had significantly lower reading scores than those who reported they almost never experience bullying ($B = -31.4$ points, $\beta = -.14$, $p < .001$). Students' gender, home resources for learning and early literacy skills remained significant predictors in this model ($p < .001$) though with slightly smaller effects than Model 1.

School factors

Model 3 incorporated factors related to the school environment to predict students' reading achievement. The factors included principal reports on school emphasis on academic success and problems with discipline at school as well as socioeconomic indicators at school such as resource shortages, school location and proportions of students classified as disadvantaged at the school. The model significantly predicted reading achievement, explaining 41% of the variance ($R^2 = .414$, $F(26,40) = 24$, $p < .001$).

Only two of the five school factors showed a significant relationship to reading achievement; school location, and school resource shortages. Students attending schools where principals reported that instruction was less affected by resource shortages were likely to score higher than those at schools where instruction was more severely affected by resource shortages ($B = 10.8$ points, $\beta = .15$, $p < .001$). School location, particularly schools outside of urban areas, had a substantial influence on reading achievement. Students attending small town or village schools and remote rural schools were likely to score between 53 and 65 points lower ($\beta = -.19$, $p < .01$ and $\beta = -28$, $p < .001$ respectively) than those attending schools in urban areas. Other school factors like proportion of disadvantaged students, emphasis on academic success and problems with discipline showed non-significant relationships with reading achievement in the model ($p > .05$).

The majority of factors from Model 2 remained significant in Model 3, indicating these factors still contribute to explaining reading achievement even after accounting for school influences.

Summary of contextual factors

These first three models demonstrate the relative influence of home, student and school factors in in determining reading achievement. While the home environment factors in Model 1 explained 23% of variance, the addition of student attitudes and engagement in Model 2 increased the variance explained to 35%. The incorporation of school variables in Model 3 improved the variance explained by a further 6%. This indicates that 41% of the variance in overall reading achievement for these students can be attributed to factors related to their contextual background. Across all three models, variables that indicated socioeconomic status either at home or of the school remained significantly associated with reading achievement. In all cases the relationship was negative, in other words, students with fewer resources for learning either at home or at school achieved lower scores overall.

Comparing individual variable results across these three models also reveals some mitigating effects. For instance, the influence of home resources for learning declines slightly in effect size, but remains significant from Model 1 to Model 3, indicating this background factor shapes achievement regardless of school quality. However, the effects of student confidence and engagement decrease in magnitude once accounting for school factors. This suggests that part of the achievement differences associated with student attitudes may be attributed to school experiences. The final two models estimated retain the contextual factors from these three models, and add factors related to the test language.

5.5.4 Modelling language factors

Test language use at home

Model 4 retained 15 contextual factors from Model 3 and added 5 predictors capturing information about students' use of the test language at home. This expanded model explained 45% of the variance in reading achievement scores ($R^2 = .447$, $F(33,31) = 15.3$, $p < .001$), a slight increase (3%) from what was explained by Model 3. This means that the addition of home language use factors explained only a small amount of additional variance in reading achievement beyond contextual factors.

When considering the individual factors associated with test language use at home there is some evidence of relationships with reading achievement at $p < .05$ level, however, the standardised coefficients show that these were relatively weak effects. Students who reported that they never speak the test language at home scored nearly 34 points lower on average ($B = -33.6$) than those who always or almost always speak the test language at home ($\beta = -.07$, $p < .01$). This relationship was not significant when considering parent reports of how frequently they use the test language to communicate with their child at home. Those whose fathers do not use the test language to communicate with them scored 14 points lower than those whose fathers do ($B = -14.46$, $\beta = -.05$, $p < .05$). However, this relationship was not significant when considering the language used by the mother to communicate with their child. Notably, students whose parents reported that they did not speak the test language at home before beginning primary school scored almost 18 points higher ($B = 17.73$) than those who did ($\beta = .06$, $p < .05$).

The influence of the other contextual factors remained largely unchanged from Model 3, with home resources for learning ($\beta = .15$, $p < .001$) and students' confidence in reading ($\beta = -.27$, $p < .001$) preserving the strongest relationships with reading achievement. School location also maintained significance as a predictor of reading achievement, with rural schools associated with substantially lower performance.

Language of the test

The final model added test language as a predictor while retaining all variables from Model 4. Model 5 explained 54% of the variance in reading achievement ($R^2 = .542$, $F(43,21) = 24.2$, $p < .001$). The inclusion of test language, therefore, contributed an additional 9% of explanatory power over and above the variance explained by contextual factors and home language use. This means that more than half of the variance in reading scores in PIRLS can be explained by the factors included in this model.

The test language itself showed a substantial relationship, with students writing the PIRLS assessment in English scoring markedly higher than other language groups. There was no significant difference for students taking the test in Afrikaans compared to those taking it in English. However, students taking the test in any of the nine African language groups all scored significantly lower than those in the English group after accounting for contextual factors and the extent to which the test language is used at home. The magnitude of the effect was marginally lower for those in the Sesotho ($\beta = -.1$, $p < .01$) and Xitsonga groups ($\beta = -.07$, $p < .01$). While students taking the test in isiNdebele, isiXhosa, isiZulu, Setswana, siSwati, and Tshivenda scored between 56 and 82 points lower on average those in the English group ($-.25 < \beta < -.04$; $p < .001$). The gap was most pronounced for students in the Sepedi group, where the average difference was over 130 points ($B = -130.2$ points, $\beta = -.36$, $p < .001$), more than 1 standard deviation on the PIRLS achievement scale.

Some home language use factors were still significant irrespective of test language, notably students who reported never using the test language at home continued to score almost 30 points lower ($B = -29.4$ points, $\beta = -.06$, $p < .01$). However, the significant positive effect seen for students who did not speak the test language before they began primary school in Model 4 was no longer significant after accounting for the different test languages ($\beta = .01$, $p > .05$). This suggests that speaking the test language at home before beginning school is different across different test language groups.

The positive effects seen for home resources for learning also continued to show a relationship with achievement, however, these effects decreased slightly in magnitude in this model (Model 4: $B = 11.7$, $\beta = .15$, $p < .001$ versus Model 5: $B = 8.5$, $\beta = .11$, $p < .001$). Notably, the most substantial change in magnitude of the effect of home resources for learning occurred between Model 2 and Model 3 once accounting for school factors (Model 2: $B = 22.1$, $\beta = .30$, $p < .001$ versus Model 3: $B = 10.36$, $\beta = .14$, $p < .001$). This suggests that the strengths of the relationship between home resources for learning and reading achievement was partially confounded by both school factors and test language.

The negative relationships related to school factors (school location, and resource shortages at school) seen in previous models also declined in magnitude after accounting for test language. Students attending remote rural schools, for instance, were likely to score 69 points lower than those in urban schools in Model 4 ($\beta = -.28$, $p < .001$). However, after including test language, the difference reduced to a 35-point gap with lower degree of magnitude ($\beta = -.14$, $p < .05$). This suggests differences in achievement related to school location are also related to the differences seen across test language and points to a potential overlap between school location and test language.

Overall, across all models the results show that there are a number of variables related to students' home, school and language backgrounds that influence their reading achievement. While home resources and confidence in reading show consistent positive relationships, and rural school location shows persistent negative impacts, the strength of many contextual factors is partially absorbed by language factors. This means that the test language explains additional achievement variation in PIRLS beyond contextual variables alone.

Table 5-4 ***Results of multiple regression models comparing coefficients and t-statistics***

	Model 1 ($R^2 = 0.23$)			Model 2 ($R^2 = 0.35$)			Model 3 ($R^2 = 0.41$)			Model 4 ($R^2 = 0.45$)			Model 5 ($R^2 = 0.54$)		
	B	**β**		**B**	**β**		**B**	**β**		**B**	**β**		**B**	**β**	
Intercept	126.51		***	243.17		***	295.34		***	302.16		***	375.21		***
CONTROL FACTOR															
Student gender															
(Girl)															
Boy	-43.33	-0.19	***	-32.96	-0.15	***	-31.51	-0.14	***	-29.3	-0.13	***	-27.08	-0.12	***
HOME FACTORS															
Home resources for learning															
	26.69	0.36	***	22.06	0.3	***	10.36	0.14	***	11.66	0.15	***	8.48	0.11	***
Early literacy activities at home															
(Often)															
Sometimes	-3.88	-0.02		-4.99	-0.02		-5.04	-0.02		-11.62	-0.05	*	-5.9	-0.03	
Never or almost never	-27.04	-0.05	*	-25.44	-0.04	*	-20.74	-0.04		-23.13	-0.04		-9.29	-0.02	
Could do early literacy tasks before school															
(Very well)															
Moderately well	-9.2	-0.04		-4.03	-0.02		-9.35	-0.04	*	-8.14	-0.04		-10.97	-0.05	*
Not well	-37.82	-0.15	***	-27.55	-0.11	***	-33.81	-0.14	***	-31.49	-0.12	***	-33.25	-0.13	***
Parents like reading															
	3.48	0.05	*	2.55	0.04		2.63	0.04		2.58	0.04		1.75	0.03	

	Model 1 (R^2 = 0.23)			Model 2 (R^2 = 0.35)			Model 3 (R^2 = 0.41)			Model 4 (R^2 = 0.45)			Model 5 (R^2 = 0.54)		
	B	**β**		**B**	**β**		**B**	**β**		**B**	**β**		**B**	**β**	
STUDENT FACTORS															
Students sense of school belonging															
(High sense of school belonging)															
Some sense of school belonging				4.75	0.02		-0.01	0		2.91	0.01		-2.49	-0.01	
Little sense of school belonging				7.17	0.02		-0.86	0		3.11	0.01		-9.39	-0.02	
Student experience of bullying															
(Almost never)															
About monthly				-11.99	-0.05	*	-2.71	-0.01		-8.27	-0.04		-5.18	-0.02	
About weekly				-31.37	-0.14	***	-27.45	-0.12	***	-31.97	-0.14	***	-27.97	-0.12	***
Student engaged in reading lessons															
(Very engaged)															
Somewhat engaged				-20.14	-0.08	**	-12.72	-0.05		-7.64	-0.03		-10.85	-0.04	
Less than engaged				-44.14	-0.09	***	-32.82	-0.07	**	-32.53	-0.06	*	-28.27	-0.05	*
Student likes reading															
(Very much like reading)															
Somewhat like reading				1.94	0.01		-5.17	-0.02		-1.23	-0.01		-7.51	-0.03	
Do not like reading				-0.44	0		-12.64	-0.03		-13.12	-0.03		-22.92	-0.06	*
Student confident in reading															
(Very confident)															
Somewhat confident				-47.76	-0.2	***	-34.09	-0.15	***	-32.77	-0.14	***	-29.36	-0.12	***
Not confident				-83.53	-0.37	***	-61.63	-0.28	***	-60.96	-0.27	***	-60.53	-0.26	***

	Model 1 ($R^2 = 0.23$)			Model 2 ($R^2 = 0.35$)			Model 3 ($R^2 = 0.41$)			Model 4 ($R^2 = 0.45$)			Model 5 ($R^2 = 0.54$)		
	B	**β**		**B**	**β**		**B**	**β**		**B**	**β**		**B**	**β**	
SCHOOL FACTORS															
Reading resource shortages															
							10.76	0.15	***	10.26	0.15	**	8.88	0.13	**
School emphasis on academic success															
							-1.11	-0.02		-1.83	-0.04		-1.52	-0.03	
School discipline															
(Hardly any problems)															
Minor problems							-5	-0.02		7.27	0.03		10.87	0.05	
Moderate to severe problems							-18.07	-0.07		-8.49	-0.03		0.09	0	
School location															
(Urban-densely populated)															
Medium size city or large town							20.23	0.04		10.83	0.02		4.32	0.01	
Small town or village							-52.66	-0.19	**	-58.96	-0.2	**	-43.58	-0.15	**
Remote rural							-64.55	-0.28	***	-69.37	-0.28	***	-34.58	-0.14	*
Suburban/township							-27.12	-0.11		-36.13	-0.15		-23.29	-0.1	
School composition by student background															
(Not disadvantaged)															
Disadvantaged							-16.92	-0.07		-19.36	-0.08		-14.52	-0.06	

	Model 1 (R^2 = 0.23)		Model 2 (R^2 = 0.35)		Model 3 (R^2 = 0.41)		Model 4 (R^2 = 0.45)			Model 5 (R^2 = 0.54)		
	B	**β**	**B**	**β**	**B**	**β**	**B**	**β**		**B**	**β**	
LANGUAGE USE FACTORS												
Test language spoken at home (student)												
(Always/almost always)												
Sometimes							13.77	0.05		5.79	0.02	
Never							-33.58	-0.07	**	-29.39	-0.06	**
Child speaks test language at home												
(Always/almost always)												
Sometimes							11.36	0.04		1.63	0.01	
Never							27.38	0.02		3.61	0	
Father speaks test language with child												
(Yes)												
No							-14.31	-0.05	*	-11.87	-0.04	*
Mother speaks test language with child												
(Yes)												
No							-5.77	-0.02		-13.23	-0.05	
Child spoke test language before beginning school												
(Yes)												
No							17.73	0.06	*	3.36	0.01	

	Model 1 ($R^2 = 0.23$)		Model 2 ($R^2 = 0.35$)		Model 3 ($R^2 = 0.41$)		Model 4 ($R^2 = 0.45$)		Model 5 ($R^2 = 0.54$)		
	B	**β**	**B**	**β**	**B**	**β**	**B**	**β**	**B**	**β**	
TEST LANGUAGE FACTOR											
Test language											
(English)											
Afrikaans									-10.96	-0.03	
isiNdebele									-64.36	-0.04	***
isiXhosa									-66.48	-0.21	***
isiZulu									-80.09	-0.25	***
Sepedi									-130.25	-0.36	***
Sesotho									-51.46	-0.1	**
Setswana									-65.18	-0.13	***
siSwati									-55.48	-0.07	***
Tshivenda									-63.05	-0.08	***
Xitsonga									-57.65	-0.07	**

Notes:
( ) reference category
B regression coefficients
β standardised regression coefficients
*Significance codes p < .001 '***'; p < .01 '**' ; p < .05 '*'*

5.6 Discussion

The disturbingly low overall reading scores, together with the substantial gaps in achievement across language groups reflected in the South African PIRLS results, underscore the importance of developing a more in-depth understanding of the factors affecting reading outcomes in South Africa. However, the diverse heterogeneity of the South African context, coupled with the extreme social inequalities mean that disentangling the factors impacting educational outcomes is a considerably challenging undertaking. This paper explored the relationships between background factors and reading comprehension for Grade 4 students in South Africa using data from PIRLS Literacy 2016. More specifically it investigated the role of contextual factors from the students' home and school environments (RQ1) in predicting overall reading achievement, alongside considering the more intricate effects related to the use of the test language at home (RQ2) and language of the test itself (RQ3).

The findings revealed that contextual factors (RQ1) related to students' backgrounds, motivation to read, home and school environments significantly predicted 41% of the variation in overall PIRLS reading achievement. The student and home factors that demonstrated the strongest relationships were gender, home resources for learning, and confidence in reading. These findings align closely with the commonly identified predictors from the international literature that suggest that factors associated with socioeconomic status, as well as student attitudes towards reading impact reading achievement (Guzmán-Simón et al., 2020; Miller & Faircloth, 2009; Nag et al., 2023; Park, 2008).

When considering school factors, the influence of socioeconomic issues on reading achievement was also evident in the findings. Substantial negative effects were seen for schools who reported that instruction was affected by reading resource shortages. Additionally, students who attended schools far-removed from major cities were significantly more likely to achieve lower results in PIRLS. The substantial negative effects seen for variables like rural school location mirror previous findings regarding inequalities in access to quality instruction in South Africa (Spaull & Pretorius, 2019). The influence of school location in South Africa is a particularly important consideration due to the ongoing impact of the apartheid regime. Historic policies of racial segregation coupled with vastly unequal government spending for schools on the basis of race, have left extreme infrastructural and resource shortages for schools in disadvantaged areas (Salisbury, 2016; Spaull & Pretorius, 2019; Taylor & Yu, 2009; van der Berg & Gustafsson, 2019). Similar findings were reflected in a detailed analysis of socioeconomic effects in PIRLS 2006 results by Taylor and Yu (2009) who found that far more variance in reading achievement in South Africa was explained by the socioeconomic status of the school than by the students' individual socioeconomic status. Their findings suggest that differences in school quality represent a critical dimension for understanding the performance of South African students.

As critiqued by Wagner (2010, 2017), findings from research in developed contexts have limited transferability to developing countries characterised by vast inequality and complex sociolinguistic dynamics. The overall low achievement and within-country disparities evident in South Africa stretch well beyond factors identified from the higher-performing settings that dominate most reading research. Added to this, the complex dynamics underpinning persistent literacy challenges across languages in South Africa mean that any discussion of reading achievement needs to consider the multilingual context carefully (DBE, 2023; Mohohlwane, 2019).

The incremental contribution of detailed home language use variables to explaining overall achievement variation was unexpectedly small (RQ2). The variables considered regarding language use related to the extent to which the test language was used at home by students, as well as by parents when communicating with their child. Accounting for factors related to frequency of using the test language at home did not make a substantial difference to the overall variance explained by the model. This finding contrasts with much of the evidence regarding the vital role of home language environment on reading development, which typically shows strong relationships especially in multilingual contexts (Dixon et al., 2012; Sénéchal & Lefevre, 2014). Therefore, there may be additional factors at play that were not accounted for in this study.

Boakye (2015) notes the vastly different literacy practices that are associated with English and Afrikaans when compared to indigenous African language groups in South Africa, as well as the profound influence these differences can have on students' reading literacy. Issues related to sociolinguistic dynamics involving cultural and socioeconomic status intertwine with reader identities and motivations in ways that are not easily disentangled (Cele, 2001; Hoff, 2013; Makalela, 2017). Moreover, close inspection of the questionnaire data revealed inconsistent patterns across different home language use measures. For instance, achievement gaps were evident when considering the students' reports of the frequency at which the test language was used at home; yet parent reports showed no significant differences.

Conflicting results such as these echo extensive critiques regarding the reliability and validity of data collection tools for assessing complex educational dynamics in developing countries (Howie, 2022; van Staden et al., 2019; Wagner & Castillo, 2014). Care should therefore be taken not to oversimplify interpretation of background questionnaires nor neglect the critical examination of limitations. Given the central role of language highlighted throughout the literature, the intricate sociolinguistic environments in South Africa warrant further investigation beyond surface-level survey measures.

The final variable added to the model sought to uncover further insight into the relationship between reading achievement and the test language (RQ3). The addition of test language to the model substantially increased the variance in reading achievement explained. This implies that the language in which students took PIRLS in South Africa has a considerable influence on their overall reading achievement. Moreover, the effect of test language supersedes the extent to which students speak the test language at home. The magnitude of the test language effect is different across language groups, with no observable effect when comparing Afrikaans students to English students, students taking PIRLS 2016 in an African language performed markedly lower than English groups, with gaps ranging from 51 to over 130 points. This means that, after accounting for the influence of contextual and home language use variables, students who take the test in, for example, Sepedi are likely to achieve a PIRLS score more than one standard deviation lower than students taking the test in English. The relationship between test language and performance seen in South Africa contrasts with more developed multilingual contexts such as Canada, where no significant differences were seen in overall scores when comparing students taking the test in French or English (Brochu et al., 2017).

Furthermore, once the language in which students took the PIRLS Literacy 2016 test in South Africa was accounted for, the effects of certain consistent contextual predictors decreased noticeably. For example, students that did not use the test language before starting school were likely to achieve higher PIRLS scores before test language was accounted for, however this effect was no longer significant after test language was considered. This pattern is further exacerbated by the test language being the medium of instruction at each school rather than the students' home language. Only 24% of students attending English schools speak the test language as a home language, compared to over 70% across the other ten groups. Despite the low proportion of home language speakers in the English group, they are the highest performing group.

A potentially influential factor here is the relationship between attending English schools and higher performance in PIRLS. The wake of historically discriminatory policies means that there continue to be substantial school quality and infrastructural differences across schools in South Africa. Typically, schools that only allowed white students under the apartheid government still use either English or Afrikaans as a medium of instruction. Although the socioeconomic and racial profile of these schools has diversified, they continue to be substantially better resourced than formerly disadvantaged schools (Salisbury, 2016; Spaull & Jansen, 2019; Taylor & Yu, 2009; van Staden & Combrinck, 2023).

Added to this is a substantial imbalance of resources, approaches, and teacher training opportunities for teaching reading in African languages. In South Africa, almost all initial teacher education programmes at universities are offered in English (or sometimes Afrikaans) (Makalela, 2016; Ngubane & Makua, 2022). Recent initiatives aim to address this by requiring trainee teachers to take an African language as a minor subject during their university degree, however, this is typically an introductory language course focused on developing a basic level of proficiency in one selected African language. Consequently, a considerable gap remains in developing and implementing approaches for teaching reading in African languages (Mohohlwane, 2019; Spaull et al., 2020). Moreover, while there are ongoing efforts to improve the availability of resources for teaching reading in African languages (e.g., books, readers, classroom materials), these materials are not yet as comprehensive or readily available as those for English and Afrikaans (McKinney, 2024; Spaull et al., 2020). This imbalanced situation means that the relationship between test language and PIRLS performance is intricately associated with current and historical socioeconomic gaps, linguistic resources, and racial inequities.

Furthermore, the negative relationship rural schools showed in initial models was reduced after accounting for test language, but nevertheless remained significant. Similarly, the significant influence of home resources for learning reduced slightly after accounting for test language but not entirely. This suggests that the effects of school location and home resources for learning remain significant even after accounting for the substantial influence of language of instruction. Such patterns suggest that language factors may act as intervening variables that help explicate how broader socioeconomic contexts influence achievement. These findings substantiate past critiques that concluding particular variables have “no impact” can be highly problematic without considering potential mediating or confounding relationships, especially in highly diverse and/or developing countries (Araújo et al., 2017; Gomendio, 2023; Heyneman & Loxley, 1983).

In South Africa, language and socioeconomic status scaffold one another, with school language policies and home language use intertwining in complex ways that ILSA data cannot easily capture. Since language represents both the medium through which the assessment takes place, and a substantial component of the construct being assessed, attempting to separate its effects from wider sociocultural contexts proves exceptionally challenging. Nonetheless, two factors – test language and school location – maintained statistically significant relationships with achievement even after accounting for their potential overlap. While the present results demonstrate concerning achievement gaps, simply controlling for language use and contextual factors fails to capture enduring barriers tied to racial marginalisation. Without insight into the specific mechanisms by which these factors exert influence, discussing “location effects” or “language effects” offers limited value for targeted intervention. These findings underscore the limitations of considering just surface-level comparisons of reading achievement and echo persistent calls for dedicated, context-specific research in developing countries.

5.7 Limitations and future research

This analysis has several limitations that should be considered carefully when drawing inferences from the findings. First, the use of self-reported background data from students, parents, and principals poses validity and reliability concerns, particularly given the low overall achievement of students. Added to this, in analysing parent and student responses to similar background questions, some concerning inconsistencies emerged. Parents tended to report higher levels of home literacy resources and activities than students reported of their own homes. Also parent and student reports of the extent to which the student spoke the test language at home did not correspond well. Moreover, the low level of reading ability shown by South African students suggests they may lack the skills, comprehension, or motivation to respond reliably and accurately to questionnaire items. The limitations of self-reported data from PIRLS questionnaires was also noted as one of the key takeaways from the recent van Staden and Combrinck (2023) book which explores the utility of ILSA data for tracking reading progress in South Africa.

Another concern emerging from the issues found with the reliability and validity of the questionnaire data is that PIRLS uses background questionnaire data to support the reliability of achievement scaling (von Davier et al., 2020). This means that when the background data does not satisfactorily represent the contextual situation in participating countries, it also compromises the reliability and validity of the achievement results. This has implications both for studies utilising the background data (such as this paper), as well as for the validity of insights that could potentially be gained from achievement results more broadly.

Furthermore, test language overlaps substantially with race and historic discrimination in South Africa. Indeed this overlap is considered to be so substantial, that a number of similar studies use test language as a proxy for race since race data is not available in PIRLS (e.g., Gustafsson, 2019; Taylor & Yu, 2009). This risks potential confounding effects related to enduring disparities tied to apartheid segregation policies. However, it is important to note that while Afrikaans and English schools historically catered to white students only, today these schools have racially diverse (but still socioeconomically advantaged) populations.

Additionally, the methodological approach involved in comparing complex models such as these risks issues with precision and power. Although efforts to moderate these issues were made through the use of replication techniques and weightings, as well as careful diagnostic checks, results should nevertheless be interpreted with caution.

Finally, the regression analysis was limited by relying on a set of pre-determined predictor variables that have been found to be influential in prior PIRLS research internationally or in the South African context specifically. While justification existed for selecting variables already shown to impact reading achievement, this inclusion risks overlooking additional factors not previously examined that may be locally relevant. Moreover, due to additional concerns about the reliability of the self-reported data, conclusions from previous research about the South African data may be unreliable. Moving forward, strengthening background data collection tools, gathering additional qualitative insights, and comparing models rigorously could help address some of these concerns.

While the limitations outlined here stress that the findings need to be interpreted cautiously, this paper's contextualised analysis offers an analytically rigorous base from which to build more nuanced investigations. The contributions extend beyond just providing a foundational understanding within the context – by explicitly grappling with the limitations of PIRLS data in the South African context, this work highlights specific areas of concern for the use of these tools in linguistically diverse, developing country settings. In doing so, it provides insights that can guide future research and instrument development that is more attuned to the complex dynamics in developing countries.

CHAPTER 6 (PAPER 2)
FIT FOR PURPOSE OR OFF THE MARK: PIRLS 2016 IN SOUTH AFRICA

Abstract

South Africa has participated in the Progress in International Reading Literacy Study (PIRLS) for more than two decades and has been consistently ranked as the lowest performing country in every cycle. Developing countries such as South Africa often rely on international large-scale assessments (ILSAs) like PIRLS to reveal insights into their education system through the lens of student performance. However, there has been considerable evidence that ILSAs fall short when it comes to assessing students from lower-performing countries accurately and reliably, which impacts the inferences that can be drawn from participation in these tests. This research evaluates whether PIRLS is able to meet the needs of the South African context, which are complex in terms of linguistic diversity, wide attainment gaps and extreme inequality. It examines whether the PIRLS Literacy 2016 instrument is appropriately targeted for students from three language groups by considering the fit of the models used, as well as the alignment between the ability of students and the difficulty of the items. The findings show that more than one third of the items did not fit the students in this study. Furthermore, more than half of the items were too difficult to provide any information for the below average students in all three groups, and only 3% of the items were appropriately targeted for below average students in the Sepedi group. Overall, the evidence suggests that PIRLS Literacy may not be a valid measure of student reading achievement in South Africa, particularly when comparing performance across language groups, and the information drawn from it may not be meaningful.

Key words

PIRLS, large-scale assessment, South Africa, targeting, item response theory

6.1 Introduction

Participation in international large-scale assessments (ILSAs) such as the Programme for International Student Assessment (PISA) and the Progress in International Reading Literacy Study (PIRLS) is usually driven by an interest in better understanding educational systems through the lens of student achievement. Countries choosing to participate in ILSAs typically do so to learn about their own and others' educational contexts for the purpose of improving educational policy and practice; but participation does not come without controversy and concerns. Reconciling the use of ILSAs to address the needs of different countries, with the technical demands of providing a high-quality assessment, is a challenging undertaking (Lockheed & Wagemaker, 2013).

One of the primary challenges is finding a way of measuring student achievement validly and fairly. Validity is a fundamental concern in large-scale assessment; it is necessary to ensure that the test accurately and fairly measures what it is aiming to measure for all participating students (AERA et al., 2014; ITC, 2001, 2019). Moreover, the importance of ensuring the validity of score interpretation is underscored by Messick (1995) due to the potential social consequences on individuals participating in assessments. There is also concern that – for low-performing countries in particular – the tests may not be appropriately targeted to the ability of the students, which affects the information provided by models and subsequently the validity of information that can be drawn from the results (Addey & Sellar, 2019; Atar et al., 2023; Rutkowski & Rutkowski, 2022; Tijmstra et al., 2020). This is a specific concern for countries where ILSAs are used to inform educational policy changes, particularly those where there is limited data available at a national level (Gomendio, 2021; Gorur & Wu, 2015). If policy decisions are made on the basis of an invalid assessment, the potential consequences may affect all stakeholders involved in the education system as well as future generations of students (Nisbet & Shaw, 2019).

In response to these concerns, and to help facilitate the inclusion of low-income and developing countries, many organisations that run ILSAs have introduced easier versions of their assessments (Addey & Sellar, 2019). However, evidence of the effectiveness of the design, equating methods and suitability of these easier versions is still lacking (Ahmed et al., 2022; Kamens & Benavot, 2011; Rutkowski & Rutkowski, 2022; Wagner & Castillo, 2014). Furthermore, although the validity and comparability of different versions of PIRLS is considered carefully by the test developers and secondary researchers at a between country level, studies considering comparability within countries are less frequent, and often less reassuringly answered (e.g., Ercikan & Por, 2020; Finch & Hernández Finch, 2016; Goodrich & Ercikan, 2019; Roux et al., 2022). This is particularly concerning for low-performing countries such as South Africa who participate to obtain detailed information about the performance of different groups within their country and are less interested in cross-country comparisons (Addey & Sellar, 2019; Atar et al., 2023; Howie, 2022; Kanjee, 2018). This paper considers the appropriateness of PIRLS 2016 for the country with the lowest average performance – South Africa. More specifically, it evaluates whether the PIRLS 2016 results are able to provide valid information about reading achievement for the purpose of comparing groups within the country.

6.1.1 What is PIRLS?

PIRLS is an international assessment conducted every five years by the International Association for the Evaluation of Educational Achievement (IEA) to assess the reading abilities of fourth-grade students in multiple countries. The study aims to gather comprehensive data on reading in various contexts, considering reading as a purpose-driven and interactive process. PIRLS includes a reading test and background questionnaires for students, parents or caregivers, teachers, school leaders, and curriculum experts. PIRLS 2016 consists of 12 reading passages with associated items, and students received booklets containing two passages and approximately 30 items. Overall scores for PIRLS are provided on a linear scale with a mean of 500 and a standard deviation of 100, a combination of advanced psychometric models and procedures are applied to provide equated and comparable overall scores for all students.

6.1.2 PIRLS 2016 in South Africa

In 2016, Grade 4 students in South Africa participated in PIRLS Literacy and achieved an average score of 320, significantly lower than the international average of 500 (Mullis et al., 2017a). This stark contrast with the international average raised concerns about the state of education in South Africa (Andersen, 2017; Govender & Hugo, 2020; Spaull & Jansen, 2019; van der Berg & Gustafsson, 2019). PIRLS establishes four benchmarks of reading achievement, and in South Africa 78% of students scored below the lowest benchmark level; indicating that a majority of students are unable to read with basic comprehension skills (Howie, Combrinck, Roux, Mokoena, et al., 2017). However, when looking at educational outcomes in South Africa, it is important to consider the profound inequalities in the context, particularly regarding language and wealth disparities as they can significantly impact reading performance (Salisbury, 2016; Spaull, 2013; van der Berg et al., 2011).

Schools in South Africa use one of the 11 official languages as a medium of instruction for the first three years of primary school, before switching to English or Afrikaans in Grade 4. Typically, this means that students learn to read in their home language, which for a majority of students is one of the nine official African languages. When comparing PIRLS results across language groups, students who took the test in English or Afrikaans had higher achievement compared to those who took it in an African language (Howie, Combrinck, Roux, Mokoena, et al., 2017). These disparities reflect the complex linguistic landscape in South African schools which still grapples with historical discrimination. A thorough understanding of these differences is crucial for addressing the ongoing educational inequities in the country.

South Africa has introduced various assessment initiatives in all 11 official languages since implementing them as media of instruction. These assessments aim to diagnose systemic challenges and track progress in closing historical achievement gaps (Zuze & Hannan, 2016). They have included the Monitoring Learning Achievement project conducted in 1999 (Chinapah et al., 2000; UNESCO, 2015), the Systemic Evaluation studies administered nationally in 2001, 2004 and 2007 (Department of Education, 2003), and the Annual National Assessments implemented each year from 2011 to 2014 (Chisholm & Wildeman, 2013; Kanjee & Moloi, 2014). Recently, large-scale evaluations like the Early Grade Reading Study (EGRS) and annual Systemic Testing have been introduced at a provincial level in selected provinces (DBE, 2022; van der Berg et al., 2022). The EGRS and Systemic Testing programme, as well as other regional studies like the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) surveys, have evaluated reading in multiple languages, but not all 11 official languages of South Africa.

This means that PIRLS is currently the only measure of reading in primary school that collects a nationally representative sample across 11 official language groups and it has become a key tool in monitoring and evaluating reading achievement at a national level (Kotzé et al., 2019; Spaull, 2022; Spaull et al., 2020; Taylor, 2017; Zuze & Hannan, 2016). PIRLS, therefore, has the potential to offer valuable insights into cross-language differences in reading performance. However, considering the low overall performance in the country compared to international comparators, it is essential to critically assess the suitability of PIRLS for the context and purpose for which it is used.

6.1.3 The purpose of participation in international large-scale assessments

In developing countries where results are typically too low for meaningful comparisons with higher-performing countries, and there is limited resource and capacity for running large-scale national assessments, participation is often driven by a desire to learn more about the situation in their own context (Addey & Sellar, 2019; Reddy, 2010; Wagner, 2010; Wagner & Castillo, 2014) South Africa is one of these countries. Despite consistently poor performance in PIRLS, as well as challenges with administration, translation and capacity; South Africa continues to pay the high costs of participation in ILSAs to gain valuable insights into their own education system and to receive the support provided for administration, data collection as well as training and development of expertise and capacity for running large-scale assessments (Howie, 2022; Kanjee, 2018; Lockheed & Wagemaker, 2013; van der Berg & Gustafsson, 2019).

The value of PIRLS to South Africa has long been debated, some common concerns include the toll testing takes on teachers and students in an already strained system; the consequences of the media's focus on poor performance compared to other countries; the relevance of the study to local policy and curriculum; the applicability of the test for the local context; the reliability and accuracy of the results; and the usefulness of the information gained from participation (Chisholm & Wildeman, 2013; Gustafsson & Taylor, 2023; Howie, 2010, 2012; Kanjee, 2018; Kanjee & Moloi, 2014; Kanjee & Sayed, 2013; Reddy, 2010; Spaull & Jansen, 2019; van der Berg et al., 2016). Despite the challenges, many researchers and policy professionals agree that a valid and reliable measure of student performance is a vital step towards redressing the historical inequalities that plague the education system (Govender & Hugo, 2020; Howie, 2022; Spaull, 2022; van der Berg & Gustafsson, 2019).

In 2019, the president announced in the state of the nation address that one of the top five development goals for the next decade in South Africa is that "every 10-year-old child will be able to read for meaning" (Ramaphosa, 2019). Focusing on reading for meaning at age 10 means that it is possible for PIRLS to be used as a method of monitoring and evaluating the progress towards this national development goal. More recently, the Minister of Basic Education stated that the PIRLS Reading Literacy Framework would be used as the basis for a shared national framework for understanding reading comprehension (Motshekga, 2021).

PIRLS, therefore, serves an important role in providing information, insight and data about students, as well as the progress and effectiveness of the education system. Owing to this and other factors, PIRLS in South Africa has grown in status and relevance over the past two decades; the results have been used to encourage policy reform, support intervention initiatives, and improve funding allocation for early grade reading in the country (Gustafsson, 2019b; Gustafsson & Taylor, 2022; Kotzé et al., 2019; Spaull, 2022; Spaull et al., 2020; van Staden & Combrinck, 2023; van Staden & Howie, 2012, 2014). While discussions around the performance in PIRLS have led South African policymakers and experts to take the important step of prioritising reading in the early grades, and in particular prioritising the teaching of reading in African languages, for this to be effective it is essential to ensure that inferences drawn from the test are valid, reliable and fair.

6.1.4 Aim and research questions

This paper aims to evaluate whether PIRLS provides valid and comparable information about reading ability across language groups that can be meaningfully used by educators and policy developers in South Africa. To do this, it examines the extent to which the models used to estimate student and item parameters for PIRLS Literacy 2016 fit the students' data in this study. Furthermore, it investigates if PIRLS Literacy 2016 is appropriately targeted for students in South Africa. In other words, it seeks to understand whether the items present a level of difficulty that is well matched to the reading ability of South African students. Three of the 11 language groups have been selected as the focus for this paper: Afrikaans, English, and Sepedi. The choice of these three languages is based primarily on the opportunities for comparison offered by the relative performance of these language groups in PIRLS Literacy 2016. The analysis addresses two research questions:

1. To what extent do the models used by PIRLS 2016 fit students' data from the Afrikaans, English, and Sepedi groups in South Africa?
2. To what extent is the difficulty level of items in PIRLS Literacy 2016 suitably targeted to the ability of students in the Afrikaans, English, and Sepedi groups in South Africa?

6.2 Literature review

6.2.1 Widening participation in international large-scale assessments

In the early iterations of ILSAs, participating countries were typically economically developed countries, primarily from Europe or North America (Peaker, 1975; UNESCO, 1962). In recent years, participation groups have widened and as a result international testing agencies have extended and tailored their assessments to suit the needs of a more diverse range of participating countries (Gomendio, 2023; Wagemaker, 2020). One of the most common changes made to facilitate this is the introduction of 'easier' versions of ILSAs. The Organization for Economic Cooperation and Development (OECD) introduced PISA for Development (PISA-D) in 2013 and the IEA, the organisation that runs PIRLS and the Trends in International Maths and Science Study (TIMSS); introduced prePIRLS in 2011, and PIRLS Literacy in 2016 and TIMSS Numeracy in 2015.

In addition to creating easier versions of the tests, the adaptions made in PIRLS have included changes to the design and development procedures, the type and extent of background data collected, as well as the analysis and modelling procedures used to determine results for countries with wide-ranging performance (Foy et al., 2018; Martin, Mullis, & Foy, 2017; Mullis et al., 2009; von Davier et al., 2023).

Making PIRLS 'easier'

In the 2011 cycle, for countries where the assessment may be too difficult for fourth-grade students, PIRLS offered two 'easier' options; countries could choose to administer a newly introduced instrument called prePIRLS or they could test older students (fifth or sixth grade) using the main PIRLS test (Mullis et al., 2012). The prePIRLS option presented a selection of easier passages and items, but the test was not equated with the main PIRLS scale. This meant that the performances of the three countries that participated in prePIRLS in 2011 (Botswana, Columbia, and South Africa) were not comparable with the 46 other participating countries in that cycle.

In 2016, the prePIRLS scale for 2011 was retrospectively rescaled to enable trend comparisons for participating countries (Martin, Mullis, & Hooper, 2017). The prePIRLS option was renamed PIRLS Literacy in 2016, which was still a less difficult option but one that was now equated with the main PIRLS test. Adaptations for PIRLS Literacy 2016 included easier passages and items, and larger font sizes as well as style more similar to texts aimed at younger readers. The most recent 2021 cycle of PIRLS did not include a separate 'easier' option, instead PIRLS 2021 applied a group adaptive design where countries can choose ratios of easy, medium, and difficult passages and items in a way that aims to present scalable results for all participants (von Davier et al., 2023).

Adapting the design of ILSAs is important because by improving and diversifying the test design, a more diverse range of countries may be encouraged to participate, and the study may be more suited to their context and needs. Adaptations are also made to the methodology, procedures and analytical techniques used to ensure the quality of the data provided by ILSAs is not compromised (Wagemaker, 2020). However, the question of whether these adapted, easier versions are still able to suit the needs and motivations of participating countries has not yet been fully demonstrated (Addey & Sellar, 2019; Engel & Rutkowski, 2021; Rutkowski & Rutkowski, 2021; Rutkowski et al., 2019).

6.2.2 Fundamentals of quality in large-scale assessments

Although much care is taken by testing agencies to ensure the design, methods and procedures used are appropriate and meet the growing demands of participating countries, assessment at such scale and range requires a careful balancing act. Validity, reliability, and fairness are three key foundational principles of quality assessments (AERA et al., 2014). However, these components are multifaceted and encompass issues related to the test features themselves as well as the context, scale, and purpose of the test in practice (Messick, 1998; Newton & Shaw, 2014). Test developers and participating countries need to reconcile the purpose of participation, the validity and reliability of the information provided, and the comparability of the information provided across diverse countries.

The constraints between what policy insights might be useful to countries, and the limitations of what ILSAs are able to (validly) provide have been raised by a number of researchers. Wagner (2010) notes that there is not a ‘one size fits all’ option and recommends that countries consider their needs carefully to balance them with the possibilities and limitations of different international assessments. Similarly, Lockheed and Wagemaker (2013) highlight the importance of aligning measurement quality with the capabilities of the assessment tools and procedures. Two of the primary factors considered essential for quality assessment across countries are the validity of the test across a wide range of countries, and the comparability of the different versions of the test in diverse contexts.

Balancing scale and validity

Although the range, scope and subject areas of ILSAs vary considerably across different tests, the application of some of the most advanced analytical models and techniques in the field of educational measurement is consistent. The most common approach across all ILSAs is to apply item response theory (IRT) models for estimating a measure of student ability on the test and a level of difficulty for each item. Central to the application of IRT models is the notion that student ability and item difficulty can be put on the same scale (Lord, 1980). This means that, in order for measurement to be valid, the difficulty of the items in the test and the ability of the students participating should be well matched (Rutkowski et al., 2019). When this does not happen, i.e., when the test is too difficult or too easy for students, there is likely to be an increase in measurement error, a reduction in accuracy, a propensity for biased scores at either end of the scale, and an overall reduction in the validity and reliability of information that can be gained (Atar et al., 2023; Zhao & Hambleton, 2017).

Recently, substantial research has underscored issues with the reliability and validity of international scores for low-performing countries (Rutkowski & Rutkowski, 2021, 2022; Rutkowski et al., 2016, 2019; Tijmstra et al., 2020). These investigations stress that despite the adaptations testing agencies have made to make their assessments more accessible to low-performing countries, there remains considerable evidence that the assessment tools are poorly targeted, and obtaining valid and reliable results for low-performing countries is challenging.

Although much of the research examines OECD studies such as PISA; concerns over the validity and reliability of IEA studies for developing countries are neither absent nor new. Gregory and Kanjee (2009) showed that the TIMSS metric was insufficient for providing a meaningful comparison between South Africa’s science achievement results and the results of other countries. Rutkowski and Rutkowski (2022) showed that TIMSS is not well suited for measuring the ability of low performing countries, particularly the seven countries that scored more than one standard deviation below the international average. Additionally, Foy et al. (2010) demonstrated that in PIRLS 2006 the lowest achieving countries had very large standard deviations, indicating imprecise achievement estimates. They noted that the PIRLS 2006 test was a poor fit for reflecting average reading levels in countries like South Africa with small proportions of correct responses as there was insufficient evidence to reliably derive proficiency scores.

In PIRLS, South Africa achieved an overall score of 253 in 2006, more than two standard deviations below the international scale centrepoint of 500 (SD 100), and lower than 300 which affects the accuracy of the measurements provided by PIRLS instruments and psychometric scales (Spaull & Pretorius, 2019). In 2011, South Africa participated in the easier prePIRLS, which was retrospectively rescaled in 2016 (Mullis et al., 2012). The procedures used for this rescaling were disputed by research in South Africa (Gustafsson, 2020; Gustafsson & Taylor, 2023). As a result, the IEA acknowledged that the originally published 2011 to 2016 trend for South Africa should not be considered reliable, however, they did not republish an updated version of the trend (Gustafsson & Taylor, 2022). This means that long-term progress trends, one of the most valuable insights that can be drawn from participation in PIRLS, are no longer available for South Africa despite having participated in every cycle since 2006. Added to this, South Africa sampled different grades for participation in prePIRLS and PIRLS in 2011 to use the PIRLS benchmarking opportunity as a way of collecting information relevant to language in education policies in the country (Mullis et al., 2012). However, in 2011 these instruments were not scaled together, and challenges with the retrospective rescaling in 2016 meant that the information sought through this dual participation option was not reliable.

In the most recent PIRLS 2021 cycle, the IEA has noted that for South Africa there are “reservations about reliability because the percentage of students with achievement too low for estimation exceeds 25%” (Mullis et al., 2023, p. 28). Similar reservations about reliability were raised for one other low-performing country, Egypt, where the percentage of students with achievement too low for estimation exceeded 15% but not 25%. However, what these reservations mean for interpreting the results, and how they affect inferences and information that can be drawn from the PIRLS 2021 data is not elaborated. These concerns for PIRLS 2021 are currently being investigated further by the National Research Committee for PIRLS in South Africa (DBE, 2023).

Balancing diversity and comparability

The issues raised regarding suitability of ILSAs concern more than the validity and reliability of achievement scores across countries. An increase in participating countries also means that the cultural, socioeconomic and linguistic domains have expanded (Wagemaker, 2020). The increased heterogeneity of participants has led to a wide range of concerns from researchers regarding the comparability of different versions of instruments related to contextual relevance, linguistic variance, and equivalence across translated versions (Arffman, 2013; Gökçe et al., 2021; Hambleton & Lee, 2013; Mtsatse & van Staden, 2021; Roux et al., 2022; Sen, 2018; Wagner & Castillo, 2014; Zhao, 2018). These comparability concerns relate to both the background questionnaires and the tests themselves. Rutkowski and Rutkowski (2022) found that there are substantial threats to the cross-cultural validity of questionnaire items, as well as concerns about the relevance of some background questionnaire items in low-income or developing contexts.

One of the primary areas of interest when discussing cross-cultural validity in ILSAs is that of language. Assessing a construct like reading ability, where language is part of the skill being measured as well as the medium through which the assessment is conducted, is a complex undertaking. This complexity is even more pertinent when the assessment is conducted in multiple language versions with the expectation of a comparable level of difficulty for students taking the test in different languages. The procedures involved in translation for PIRLS are carefully designed and regulated to consider the factors that might affect cross-language and cross-country comparability (Martin, Mullis, & Hooper, 2017; Rutkowski et al., 2013; van de Vijver & Tanzer, 2004). However, despite the rigorous undertaking of developers, several studies have pointed to a lack of comparability across different language versions of PIRLS both across and within countries (Goodrich & Ercikan, 2019; Grisay et al., 2009; Sandilands et al., 2013; Stiff et al., 2023; Stubbe, 2011).

In South Africa, researchers looking at the comparability of PIRLS across language versions found evidence that up to a quarter of items present varying levels of difficulty for students from different groups (Mtsatse & van Staden, 2021; Roux et al., 2022). Combrinck (2020) looked at issues of comparability of trend items by analysing the isiZulu versions of PIRLS in both 2006 and 2016 and found that around 20% of the items had different levels of difficulty across the two cycles. Although a level of difference across versions is to be expected, and perfect equivalence is acknowledged to be impossible to achieve, findings such as this support the ongoing debates regarding whether various ILSAs do validly, reliably, and fairly assess students across different groups. This is even more relevant in South Africa as the achievement of students from different language groups varies substantially (Howie, Combrinck, Roux, Mokoena, et al., 2017).

6.3 Methodology

6.3.1 Data

The PIRLS Literacy 2016 national dataset for South Africa (N=12,810) was retrieved from the IEA international data repository using the EdSurvey package in R (Bailey et al., 2020). A subset of 4,215 students was then selected based on the language in which students were tested. The English group had the most students (2,089) followed by Afrikaans (1,228) and then Sepedi (898). Although the sample is nationally representative, a higher number of English classes (and by extension students) were selected due to implicit stratification in the sampling procedures (English = 57 classes; Afrikaans = 38 classes; Sepedi = 21 classes). The item set for analysis includes 183 items across 12 passages; 91 multiple choice items and 92 constructed response items; 26 polytomous (0, 1, 2 or 3), and 157 dichotomous (0, 1).

6.3.2 *Item response theory models*

The study applies methods from an IRT framework, the most common method used for the design and analysis of ILSAs. IRT is a collection of measurement models that describe the interaction between the students and items by determining the probability (*P*) of a student (*j*) getting a particular item (*i*) correct based on their overall ability on the test (θ). There are variations, constraints and additional parameters that can be applied to the interaction between student ability (θ) and item difficulty (*d*) which manifest in distinct models of IRT. In this way, IRT models can provide insight into student achievement on a test, the items on the test, and the interaction between student achievement and the items on the test.

Evaluating the applicability of a test for a group of students, therefore, depends on the ability of the IRT models used to accurately reflect the distance between the ability of the students taking the test and difficultly of the items in the test. For this analysis, consistent with the approach used by IEA for PIRLS 2016 internationally, a 2-parameter logistic (2PL) model was used to estimate parameters for dichotomous items, and a generalised partial credit model (GPCM) was used for polytomous items. Alongside information about student ability and item difficulty, the 2PL model and GPCM enable information to be provided on the relative item discrimination (*a*) for each item. The inclusion of the individual item discrimination parameter (a_i) means that the discrimination (or slope) of the items can be estimated for each item, rather than for the test as a whole. This enables information to be provided on the relative item discrimination as it may vary by ability and item difficulty. The equations that follow describe each of the models, in each of the equations:

i is any specific item on the test
j is any student taking the test
P is the probability of a correct response
P_{ij} represents the probability that student *j* answers item *i* correctly based on the interaction between the following parameters:
θ – person ability estimate. It is a value that is calculated based on the students' total performance on the test itself.

a – item discrimination parameter. This describes the level of discrimination provided by each item.

d – item difficulty parameter. This represents the point at which students with average ability (i.e., θ = 0) have a 50% chance of getting a correct response on an item.

Equation 1: 2-Parameter Logistic IRT Model

$$P_{ij} = \frac{\exp\left[a_i\left(\theta_j - d_i\right)\right]}{1 + \exp\left[a_i\left(\theta_j - d_i\right)\right]}$$

The GPCM expands from the 2PL model to include items that are polytomous (Muraki, 1997). The GPCM, developed by Muraki (1992), is an extension of the partial credit model (PCM) originally proposed by Masters (1982). The primary difference between the PCM and the GPCM is that in the generalised model, the item discrimination parameter (*a*) is able to vary across items. Along with modelling the probability of a correct response for a student on a particular item, the GPCM also seeks to model the probability of scoring in a particular category, i.e., the probability that student *j* on item *i* is likely to score either 0 or 1 or 2.

Equation 2: Generalised Partial Credit IRT Model

$$P_{ik}(\theta) = \frac{\exp\sum_{j=0}^{k} 1.7a_i\left(\theta - d_i + t_j\right)}{\sum_{i=0}^{m-1} \exp\sum_{j=0}^{i} 1.7a_i\left(\theta - d_i + t_j\right)}$$

P_{ik} is the probability of responding in category *k* (*k* = 1, 2, …, *m*) on item *i*

k – response category

m – number of response categories (in PIRLS Literacy 2016 *m* = 3 or 4)

a_i – item discrimination parameter

d_i – item difficulty parameter

t_j – deviation (threshold) parameter for a student (*j*) on an item. It defines how far from the item difficulty (d_i) the threshold is located.

In GPCM, *k* represents the category of the response and *m* represents the number of possible response categories for an item. In the case of PIRLS Literacy 2016 the number of categories for all polytomous items is either three (incorrect, partially correct, or fully correct) or four (incorrect, partially correct, almost correct, or fully correct). An additional parameter that forms part of the GPCM is the deviation or threshold parameter (*t*) which represents the location at which the probability of a response in two adjacent categories is equal. In other words, *t* represents the point at which a student has an equal chance of either scoring 0 or 1 on an item, or of scoring 1 or 2 on an item.

Assumptions of item response theory models

Applying an IRT model to assessment data such as this rests on two underlying assumptions, (i) that the construct being tested is unidimensional, i.e., only one construct is being assessed; and (ii) that the items do not show local dependence, i.e., the answer to one item is not dependent on the answer to another. Unidimensionality is an assumption that is rarely met in actual testing situations, particularly in tests with common stimulus materials (e.g., passages in PIRLS) (Andrich & Marais, 2019). The dimensionality of a test is therefore often considered in terms of ensuring it has a dominant construct, rather than only one construct.

Local independence relies on items being independent of each other and is also particularly difficult to defend in reading assessments such as PIRLS Literacy 2016 where items are all related to one stimulus passage. Violating the assumptions of IRT models can have consequences for both the scale and the reliability of the estimates provided by the models (Andrich & Marais, 2019).

6.4 Analysis

In this analysis the fit of the IRT model to the data was evaluated, the underlying assumptions of the IRT model were checked to ensure there were no violations, and finally the extent to which the difficulty of the items in PIRLS Literacy 2016 suitably targeted the ability of the students was examined. The first stage of the psychometric evaluation in this study involves fitting the 2PL model for dichotomous items and GPCM for polytomous items to the data. However, before these models can be estimated some preparation and recoding of the data was necessary (a description of the steps taken in data preparation and recoding is presented in Appendix B). All IRT analyses were conducted using the mirt package version 1.4 in R which applies the marginal maximum likelihood method with an expectation-maximum algorithm to estimate models (Chalmers, 2012).

6.4.1 *Evaluating model fit*

Evaluating the overall fit of the model to the data is an essential step because fundamental issues of model suitability are in question due to the low overall performance of South African students (Atar et al., 2023). Furthermore, because a smaller subset of student response data (compared to the primary set) is being analysed there is the potential that issues of fit may be introduced (Atar et al., 2023; van Rijn et al., 2016; Zhao & Hambleton, 2017). As recommended by Chalmers (2018) the fit of the data to the IRT model was evaluated using comparisons of Akaike Information Criterion (AIC) and the Bayesian Information Criteria (BIC) (which accounts for additional parameters in the model), as well as a likelihood ratio test (χ^2). The AIC and BIC do not provide information on the model fit in isolation, but rather, by comparing the AIC and BIC values of a set of competing models for the same data – the best fitting model can be identified (Burnham & Anderson, 2002).

6.4.2 *Checking model assumptions*

Unidimensionality of the items was analysed using an Exploratory Factor Analysis technique known as Item Factor Analysis (IFA). IFA involves allowing the items to load on all possible factors estimated for both a unidimensional model and a multidimensional model and then comparing the factor loadings for the items in each model (Mair, 2018). Factor loadings are an index of the strength of the relationship between the item and an underlying factor (in this case the underlying factor is the construct, i.e., reading ability). Given that the items in PIRLS Literacy 2016 are presumed to involve correlated factors, as the test is designed to assess a single construct, the oblimin factor rotation method was applied when estimating factor loadings for the multidimensional model (Chalmers, 2012; Yong & Pearce, 2013).

Local independence was investigated by inspecting potential correlations between items using Yen's Q_3 statistic, which gives a measure of the correlation between item residuals, thereby flagging any unwanted correlation between item responses that should function independently (Yen, 1984). The dependency between items was investigated using Yen's Q_3 statistics, and items were flagged as locally dependent if $Q_3 > .20$, a conservative cut-off recommended by Andrich and Marais (2019) and Chen and Thissen (1997).

6.4.3 Examining targeting

Targeting considers the relationship between the students and the item parameters as estimated by the model. For a test to effectively assess a range of student abilities, some items are expected to have a high proportion of correct responses (i.e., easy items), while others are expected to have a low number of correct responses (i.e., difficult items). However, in assessments where there is a potential flooring effect, i.e., a high number of items have the potential to be too difficult for most participants (as is the case in this dataset), understanding if there were an unexpectedly high number of items that had a low proportion of correct answers is a vital consideration (Hessling et al., 2004). The proportion of correct responses for items should not be considered in isolation; and the graphical inspection of the student and item parameters is recommended as a means of providing a comprehensive picture of the overall alignment between students and items in the test (Hambleton et al., 1991; Kalinowski, 2019; Wilson, 2009).

The fit of items to the model was inspected by comparing the expected curve of the model in terms of the item response functions (IRF) for each item, and the observed responses along the same scale. When modelling using IRT models, the IRF provides an expected curve of the probability that students with varying levels of ability are able to respond correctly to an item. A well-functioning item should show a match between the likelihood that students respond correctly to the item ($P(\theta)$), and the overall ability of the students (θ). It is also possible that items might show reasonable fit overall, but misfit for a specific group – particularly considering the differential performance of the groups in this study so observed and expected IRFs for each group were considered with specific attention to differences in discrimination parameters for each item.

For items to be considered well-discriminating for all levels of student ability, higher item discrimination parameters (close to 1) are desirable (Meade & Wright, 2012). Items with low discrimination capacity (i.e., $a < .50$) were flagged. Additionally, the alignment between the distribution of all items in terms of their difficulty and the distribution of all students in terms of their ability was inspected using a Wright Map, which enables the graphical examination of the relationship between students and items by plotting both items and students along the same scale (Wright & Stone, 1979).

6.5 Results

6.5.1 Model fit

The comparative evaluation of a simple IRT model (Rasch model) and mixed IRT model (2PL model for dichotomous items and a GPCM for polytomous items) found that the mixed IRT model was the best fitting model as both the AIC and BIC values were lower than for the simple IRT model, see Table 6-1. The results of the likelihood ratio test (χ^2) show that the mixed model has significantly better fit than the simple model at the p < .001 level (χ^2 (182) = 2967.8, p < .001). This means that the mixed IRT model was the best model to describe the dataset.

Table 6-1

Comparison of fit statistics for simple and mixed IRT models

Model	Dimensions	AIC	BIC	χ^2	df	p-value
Simple (Rasch)	1	145497.6	146849.4	-	-	-
Mixed (2PL & GPCM)	1	142893.8	145400.7	2967.8	182	< .001

The relative fit of uni- and multidimensional models were then compared to evaluate whether there is evidence of more than one construct being assessed, as shown in Table 6-2. The lower BIC values, which penalise for additional parameters in the model, suggested that a unidimensional model fit better, however, the likelihood-ratio test and AIC values indicated better fit of the multidimensional model. These conflicting findings demonstrate that there are potential signs of multidimensionality in the model which can occur when the test is measuring a secondary trait that is either an aspect of the same construct or related to a different construct. The next stage explores these indications of dimensionality of the items in more depth.

Table 6-2

Comparison of fit statistics for uni- and multidimensional IRT models

Model	Dimensions	AIC	BIC	χ^2	df	p-value
Unidimensional	1	142893.9	145400.7	-	-	-
Multidimensional	2[9]	142600.7	146262.6	657.1	182	< .001

6.5.2 Dimensionality

Item-level dimensionality was investigated using IFA; items were flagged if they showed evidence of no relationship with a single factor in the unidimensional model (i.e., F1 < .32); if they showed a good relationship with a second factor in the multidimensional model (i.e., F2 > .63), or if they showed evidence of cross-loadings in the multidimensional model (i.e., both F1 and F2 > .45). These relatively stringent cut-off values, suggested by Tabachnick et al. (2007), were considered adequate for this dataset as the sample size was sufficient (more than 10 responses per item) (Costello & Osborne, 2005).

Although most items loaded well on a single dimension, the IFA analysis flagged dimensionality concerns for 15% of the items (28 items). Table 6-3 shows 11 items that were flagged as misfitting because they showed no relationship with the primary construct (F1 < .32), and/or a poor relationship (F1 and F2 < .45) with a potential secondary construct. Item14 did show some

[9] It is also possible that further dimensions might provide a better fit, so a three-dimensional, and four-dimensional model were also estimated. However, the addition of more dimensions resulted in the models failing the recommended convergence tests, impacting the reliability of these models. Moreover, the model fit comparisons between the unidimensional model and all three versions of the multidimensional model (two-, three-, and four-dimensions) were very similar.

evidence of cross-loadings (F1 and F2 > .32 and < .45) in the multidimensional model, however, the low factor loadings show that this is still indicative of a poor relationship with the construct. This means that these 11 items were not able to measure the same construct as the other items in the test.

Table 6-3

Misfitting items in both uni- and multidimensional models

Item	Unidimensional model	Multidimensional model	
	F1	**F1**	**F2**
Item14	0.22	0.35	0.34
Item22	0.20	0.17	-0.09
Item45	0.14	0.13	-0.04
Item66	0.31	0.29	-0.08
Item70	0.30	0.31	0.01
Item79	0.28	-0.03	0.27
Item85	0.15	0.15	-0.01
Item86	0.22	0.22	-0.01
Item87	0.30	0.28	-0.08
Item89	0.32	0.31	-0.02
Item90	0.14	0.08	-0.21

Table 6-4 shows the other 17 items that were flagged as misfitting as they showed a good relationship with a secondary factor. Unlike the items that showed no relationship with either factor in Table 6-3, these items show evidence that they are measuring a secondary construct and not the dominant construct being measured by other items in the test (i.e., F1 < .32 and F2 > .63). It is possible that these items do not necessarily violate the assumption of unidimensionality. Instead, they may indicate the existence of a sub-dimension related to reading that is still included within the main construct accounted for in the unidimensional model. Alternatively, there might be unwanted connections between the items, leading to the presence of another dimension. The next stage in this analysis, therefore, investigates whether there are dependencies found between items that violate the assumption of local independence.

Table 6-4

Items related to secondary factors in multidimensional model

	Multidimensional model	
Item	**F1**	**F2**
Item34	0.07	0.73
Item36	0.11	0.72
Item39	-0.14	0.70
Item40	-0.04	0.79
Item78	-0.01	0.71
Item80	-0.04	0.78
Item82	-0.13	0.65
Item116	0.09	0.68
Item117	0.05	0.70
Item118	-0.14	0.71
Item149	0.25	0.90
Item150	0.32	0.93
Item151	0.00	0.82
Item152	-0.30	0.76
Item153	-0.13	0.77
Item154	-0.17	0.75
Item175	-0.12	0.66

6.5.3 Dependence

Five item pairs were flagged as locally dependent at the $Q_3 > .20$ level, shown alongside their passages in Table 6-5. Interpreting the potential violations of local independence is a complex task, particularly when there is not an apparent conceptual or structural link between items. Considering that three of these pairs involved dependencies across different passages, the overall effect of their dependencies may be negligible. However, the two pairs where items occurred in the same passages also showed a relationship with a secondary factor in the dimensionality checks; revealing that these items may be assessing something other than reading ability. Further investigation is therefore necessary to understand if these potential violations are a cause for concern in other aspects of fit, and what the impact of using this model means for reliability and validity of the results. The next stage of the analysis uses the item and student parameters estimated by the model and considers their alignment to understand the extent to which the results can be meaningfully interpreted.

Table 6-5

Item pairs flagged for local dependence

Pair	Passage Title	Items	Q_3
1	The Pearl	Item116	0.22
		Item117	
2	African Rhinos and Oxpecker Birds	Item149	0.22
		Item150	
3	Hungry Plant	Item49	0.23
	Flowers on the Roof	Item181	
4	Bhagita's Perfect Orange	Item14	0.23
	Sharks	Item156	
5	Bhagita's Perfect Orange	Item16	0.26
	How Did we Learn to Fly	Item65	

6.5.4 Targeting

Proportion of correct responses

Figure 6-1 shows the number of items by their percentage of correct responses for all students in the dataset as well as by group. Items with more than 60% correct responses fall at the easier end of the distribution of item difficulty, while those with fewer than 40% correct responses can be considered more difficult items for these students. The range of easy to difficult items in Figure 6-1 is determined by the proportion of correct responses from students in this dataset. When considering all groups, 37 items (out of 183) fell at the easy end of the scale for all students, and 65 items were at the more difficult end of the scale. The disproportionate level of difficult items is even more stark when considering the results across groups, while the Afrikaans and English groups have a reasonably comparable spread of easy or difficult items. For the Sepedi group almost all items (134 of 183) fall at the more difficult end of the scale, with no items at the easy end. In the Sepedi group, just 11 items were answered correctly by more than 60% of students. This means that the vast majority of items (172 out of 183) are too difficult for almost half of the Sepedi students.

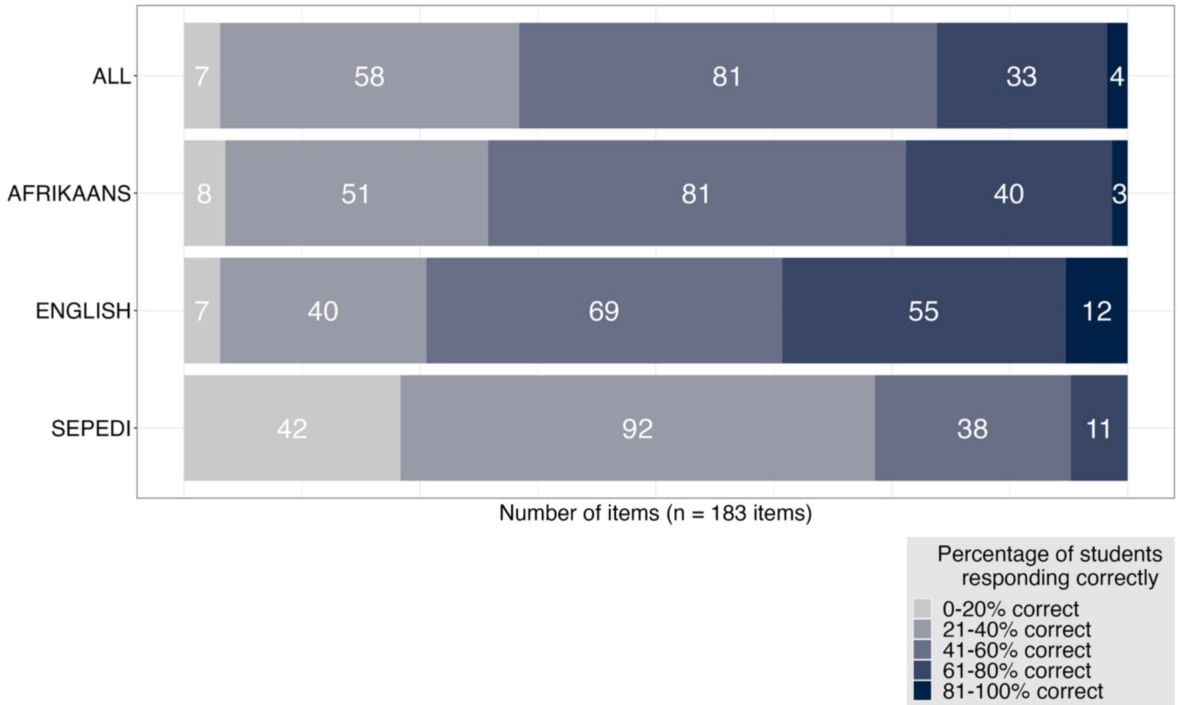


Figure 6-1 Percentage of students responding correctly to items in each group

Accurate item and student parameter estimation is dependent on sufficient data being present for all groups across all response categories. This is because items with very few correct responses can impact the reliability of both item and student parameter estimation and lead to an over inflation of item difficulty estimates, and biased student parameter estimation (Svetina et al., 2013). This is a fundamental concern (and indeed a leading motivation for conducting this analysis of PIRLS Literacy 2016 in South Africa) as having a high number of items that most students are unable to answer correctly means that the test as a whole is not able to provide information about students' reading ability. However, although this concern remains central to the research aim, at this stage of the analysis it is important to remove these low-response items so that the subsequent checks can be conducted with model parameter estimates that are as reliable as possible.

Analysists typically recommend eliminating items with fewer than 30 responses in a category overall, and no less than 10 responses per group in a category (Dai et al., 2021; Linacre, 2002). One of the challenges with eliminating items, however, is that deleting many items can have implications for the validity of the assessment from a construct coverage perspective (El Masri & Andrich, 2020). In this study, therefore, in an effort to optimise parameter estimation as much as possible and avoid removing items in a way that compromised construct coverage, items were eliminated from subsequent analyses if they had fewer than 15 correct responses per group (14 items in total). In all 14 items, the low responses occurred in the Sepedi group, and in two cases they occurred in both the Afrikaans and Sepedi group. In addition, some polytomous items required 'partially correct' categories to be collapsed to ensure that there was sufficient data (> 15 responses) in each response category. It is notable that no items had low responses in the 'incorrect' category; this means that no items in this test could be considered 'too easy' for these students.

The reduced item set that is used for the next stages of this analysis includes 169 items; 90 multiple choice items and 79 constructed response items; 8 polytomous (0, 1, 2), and 71 dichotomous (0,1). To ensure that the reduction of the item set did not affect the results of the model fit comparisons conducted previously, model fit was re-evaluated. Upon re-evaluation, the complex model remained the best fitting model at a $p < .001$ level (χ^2 (168) = 2938.5, $p < .001$), and the AIC and BIC values of the complex model remained lower than those of the simple model.

The 'ideal' number of easy or difficult items for a test is not always straightforward to determine, this is particularly true for a test like PIRLS Literacy 2016 where, due to a matrix block administration procedure, each student only responds to ~30 (of 183) items (for more information on the booklet administration procedures see Appendix A-2). Added to this, because the spread of item difficulty is designed for the full distribution of student abilities internationally, and South Africa performed at the bottom end of the scale, it is expected that there would be a higher number of difficult items for this dataset. This means that considering proportions of correct responses, while informative, is insufficient to gain a full understanding of the relationship between items and students in this dataset. The fit of each item to the model, as well as the alignment between the item difficulty estimates and the student ability estimates from the IRT model is therefore also examined graphically.

Item response functions

IRFs for each item were inspected and where there was not a strong relationship between the expected curve (grey line) and observed responses (dots with red error bars) items were flagged as misfitting. Each dot represents the observed responses of 15 students in the dataset. The graphical inspection of IRFs revealed 21 items that showed misfit between the model expectations and student responses at an overall level. Figure 6-2 shows examples of an IRF showing good fit (Item101) and one showing misfit (Item85) (IRFs for all 21 misfitting items can be found in Appendix D-1).

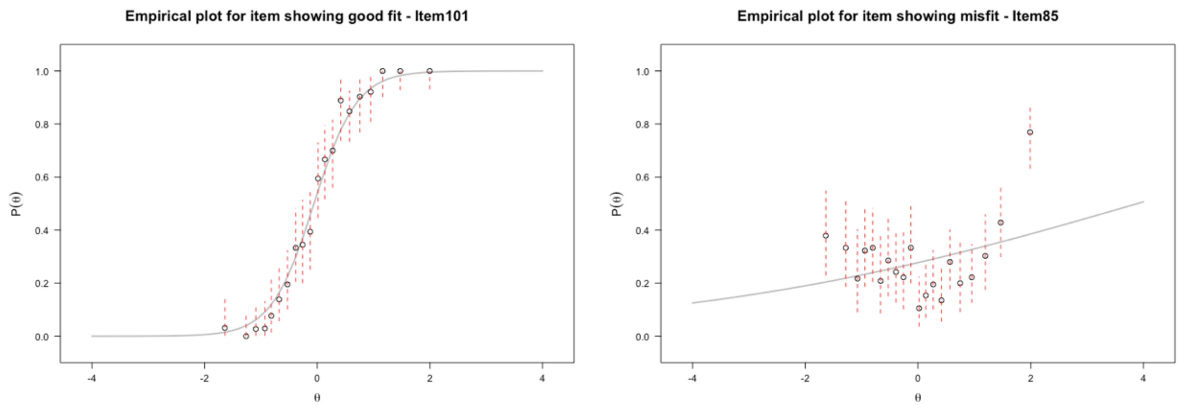


Figure 6-2 Expected versus observed item response functions – examples of good and poor fit

Following the IRF checks, items with low discrimination parameters were then flagged. The results across groups are shown in Table 6-6. Eight items in the Afrikaans group, and five in the English group had low discrimination values; while in the Sepedi group, 48 items had discrimination values below .50. To understand potential reasons behind these low discriminating items, and how they impact comparability of the items, the expected versus observed IRFs for discrimination values below 0 were examined across groups, two examples are shown in Figure 6-3. It is evident through these comparisons, that there are clear signs of misfitting items for the Sepedi group predominantly as there is no clear pattern of higher performing students having a higher probability of answering questions correctly. Twelve items discriminated negatively for the Sepedi group (i.e., $a < 0$), while two discriminated negatively for Afrikaans, and none for English. This means that, rather than finding an item easier as the student ability increases, for students from the Sepedi group the items were more difficult for students with higher ability levels.

Table 6-6

Count of poorly discriminating items by group

Group	Negative ($a < 0$)	Low ($0 > a < .50$)	Total ($a < .50$)
Afrikaans	2	6	8
English	0	5	5
Sepedi	12	36	48

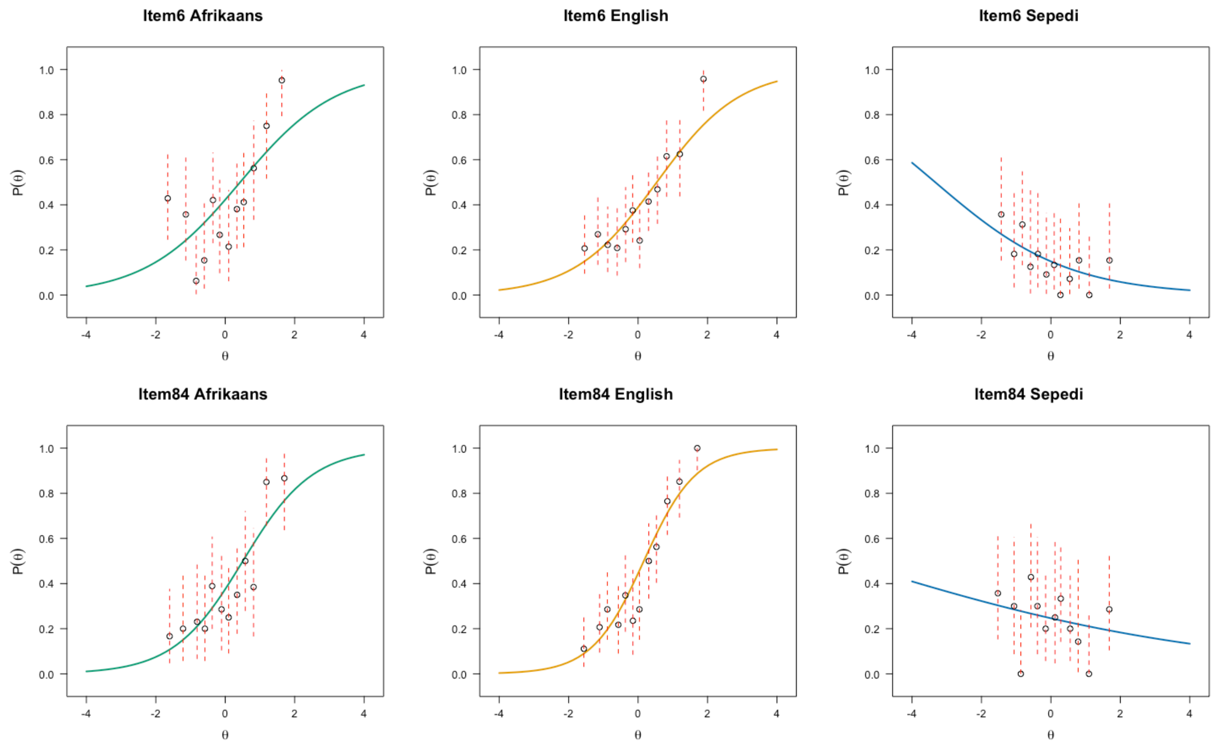


Figure 6-3 Expected versus observed item response functions – examples of negatively discriminating items across groups

Examining the observed responses for all groups did reveal misfit in the Afrikaans and English groups, although not as severe or frequent as seen in the Sepedi group. Two items (Items 14 and 85) for the Afrikaans group and two for the English group (Items 85 and 90), showed less than adequate alignment between the observed and expected curves (IRFs across groups for all negatively discriminating items are available in Appendix D-2). In 37 items, students had a higher-than-chance probability of a correct answer for multiple choice items. These items (i.e., $a < .25$) were also imbalanced across groups, one item discriminated lower than chance for the English students, six for Afrikaans students and 27 for Sepedi students. This means that students from the Sepedi group were more likely guessing the correct answer in a third of the multiple-choice questions.

Wright maps

The alignment between items and students was then examined using a Wright Map, shown in Figure 6-4. The histogram on the left represents the distribution of student ability for this dataset, on the right the items are ranked according to the item difficulty. Most students fall between $-2 < \theta < 2$ along the ability scale, and a majority of the items occur between these two points indicating that the items were somewhat well targeted for the sample ungrouped (N=4,215). However, the sparseness of items at the lower end of the ability scale, together with the positive skew in the student distribution, indicates that the test is not well targeted for these students, particularly those at the lower end of the ability scale. Indeed, students with ability estimates below $\theta < -2$ are targeted by no items at all. While nine items towards the top of the distribution (indicated in the red block) were too difficult for all students.

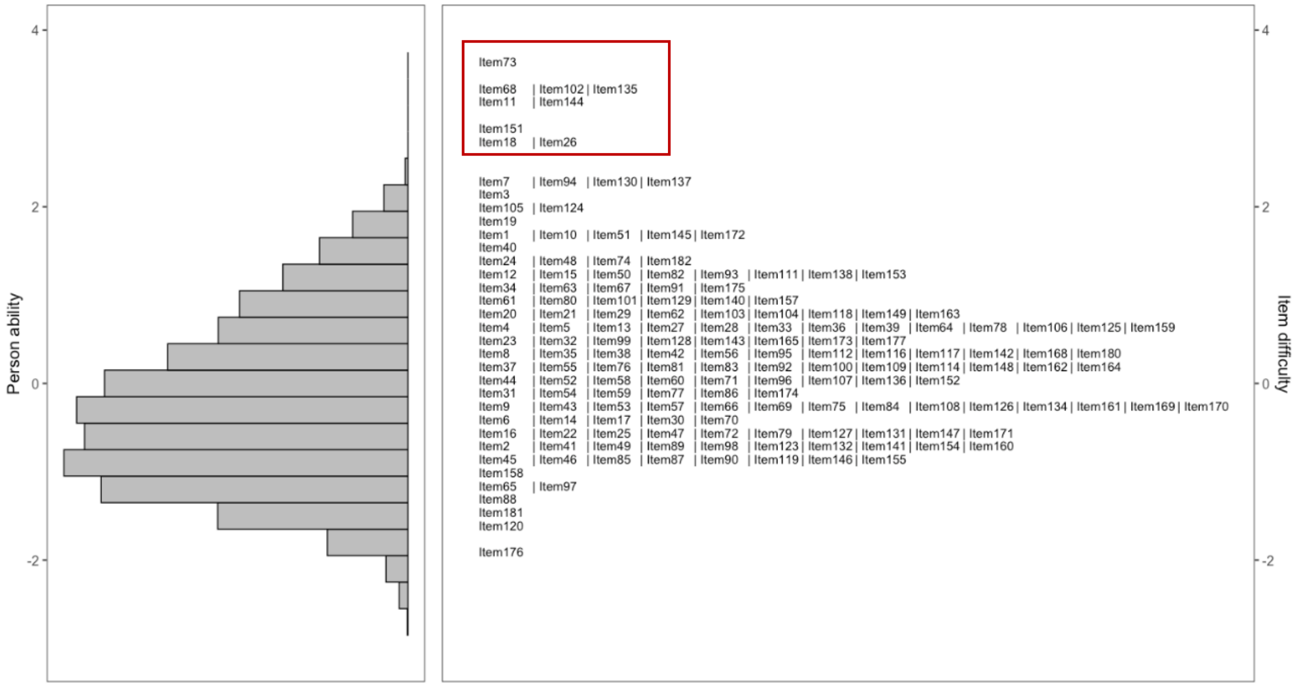


Figure 6-4 Wright Map showing distribution of student ability and distribution of item difficulty

Due to the dissimilarity between student ability distributions across language groups; targeting was also considered for each group separately (see Figure 6-5). For Sepedi students, ability is distributed between -2 < θ < .55, and both the English and Afrikaans groups are distributed between -2.5 < θ < 2.5. The annotations in Figure 6-5 highlight the items identified as having poor targeting for specific groups. For the English group just nine items fall outside of their ability range (green block), for the Afrikaans group which has a slightly narrower distribution 14 items are too difficult to achieve (yellow block), while almost a third of the items (55) are too difficult for all Sepedi students (blue block).

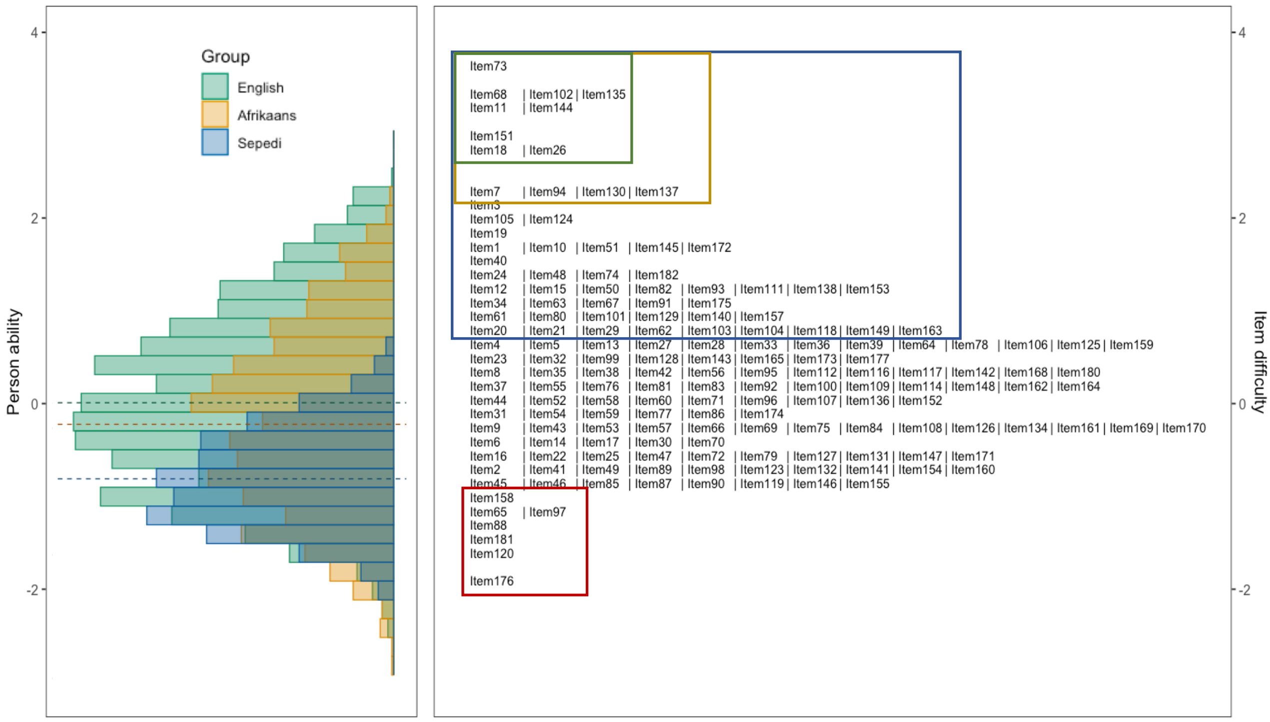


Figure 6-5 Annotated Wright Map highlighting misaligned items for all three groups

Another important consideration is the balance of items targeting students with below and above average achievement in each group. The mismatch between the items and students in this dataset means that students at the lower end of the distribution in each group are targeted by fewer items. This is particularly true for the Sepedi group, where a vast majority of students (>90%) have ability estimates below the overall average for all students in this dataset. Ideally, it would be expected that even numbers of items target below average and above average students. However, in this dataset 69 items fall below the mean versus 100 items that occur above. This means that a majority of the Sepedi students, and about half of the English and Afrikaans students, are targeted by less than half of the items. Moreover, only seven items (red block) are achievable by almost half of the Sepedi students ($\theta < -1$). This imbalance suggests that there is little statistical information available from this test for most of the students in this dataset, and almost all of the students in the Sepedi group.

These challenges with information provided due to poor alignment between students and items are further amplified by the matrix block design of PIRLS Literacy 2016. Although there are 183 items, these are distributed across 12 passages, and each student only responds to two passages with approximately 15 items in each. Therefore, for lower-performing students, where very few items are aligned with their ability, there is a relatively high chance that they only see a small percentage of the items that target them. In the case of lower-performing Sepedi students, the seven items that match their ability occur across six different passages, this means that it is highly unlikely that half of the Sepedi students received even one item that suits their ability.

6.6 Discussion

There have been changes in the public and academic conversations around results from ILSAs in South Africa over the past two decades. In many cases, reports both in the media and in research circles do not refer to South Africa's ranking in PIRLS relative to other participating countries, but rather to inequalities and differences in performance across groups within the country. However, although this within-country focus reduces some concern about the negative consequences of concentrating on international rankings, by neglecting the distance between South Africa's performance and the performance of other participating countries, the consequences that those vast differences in performance may have for the validity and reliability of the instruments are also neglected.

Recent research into the results and methods used in ILSAs to quantify the achievement of increasingly diverse ranges of student performance has shown that wide gaps between achievement in countries can lead to biased, unreliable, and invalid results (Ahmed et al., 2022; Rutkowski & Rutkowski, 2021; Rutkowski et al., 2019; Tijmstra et al., 2020). Considering the status and influence of PIRLS results in South Africa, the impact of the relatively low performance on the reliability and validity of the findings needs to be examined carefully.

Key findings from this analysis revealed that more than one third of items (35%) in PIRLS Literacy 2016 were flagged as misfitting for all students in the dataset due to concerns of fit and alignment. Fundamental IRT model assumptions of local dependency and unidimensionality were violated. Although the dependencies were minor, the analysis of dimensionality suggested that a number of items did not show any relationship with the construct. Additionally, multidimensional items were found, suggesting that some items are assessing a construct other than reading ability for these students and raising concern about the overall accuracy of the model estimation. Violating unidimensionality can lead to a distortion of item parameters, overestimation of overall scores, and underestimation of measurement errors (Andrich & Marais, 2019; Rajlic, 2019). It is important to note that although careful consideration of model fit is recognised as fundamental, the consequences of poor fit are often found to be negligible, and many researchers agree that achieving perfect fit is unachievable (Box, 1976; Reise, 1990; van Rijn et al., 2016). When an instrument fails to fit, however, the consequences to the validity and reliability of the inferences that can be drawn from the model can be substantial (Zhao & Hambleton, 2017). The more critical question therefore is whether the instrument is fit for purpose (Newton, 2007).

South Africa's participation in PIRLS is motivated by desires to use the information from PIRLS to help design and implement a strong, focused and well-designed plan to improve reading that helps to redress inequality seen in the results of PIRLS across groups (Spaull, 2022). One of the primary findings of this study is that a large number of items in PIRLS Literacy 2016 were too difficult for students in South Africa, meaning that almost no students were able to answer them correctly. This was particularly true for Sepedi students. Having difficult items in a test is useful as they help us to discriminate between high, average, and low-performing students. Although the low performance of South African students compared to other countries means that it is expected that there would be a high number of difficult items; having many items that are unachievable for almost an entire group of students means that further information about the reading ability of those students is limited.

Items with low responses in any category (correct or incorrect) also influence the estimation of model parameters. Some of these effects were seen in the inspection of item fit; a number of items were not able to discriminate between high and low ability students. This was specifically the case for Sepedi students, where the probability of a correct response for average-ability students on some items was worse than chance. A similar finding was shown in an analysis of PISA-D, in eight of the nine countries that participated, average-ability students had a lower probability of getting a correct answer than if they just guessed from the four multiple choice options (Rutkowski & Rutkowski, 2021). The validity of an instrument relies on student responses showing a relationship with model expectations to provide information about student ability. However, when chance is a more accurate predictor of correct answers than student ability, it is clear that the items are not providing sufficient information about the reading ability of those students. This has an influence on both the reliability and validity of the instrument for these students, and as a result the information that can be learned from the test is compromised.

The alignment between the distribution of item difficulty and the distributions of student ability revealed that, for the dataset as a whole, nine items were too difficult for all students. This is expected given the low overall performance of South Africa. However, when considering the alignment by group, there is a much less balanced picture, particularly for the lower performers. Below average students in the Afrikaans and English groups are targeted by approximately one third of the items, whereas low-performing Sepedi students were targeted by just seven items. Similar problems were seen when looking at below average students from low-performing countries in PISA. For the Dominican Republic in PISA 2015 just two items targeted below average students (Rutkowski et al., 2019); while below average students from Zambia were measured by no items at all in PISA-D (Rutkowski & Rutkowski, 2021).

Participation in PIRLS is costly, and the opportunities it offers to draw detailed insights into reading achievement, classroom practice, home and school context and student attitudes towards reading are a leading motivation for South Africa's participation. Additionally, the extensive and rigorous procedures involved in translating PIRLS into 11 official languages provide a rare opportunity to explore the effectiveness of the multilingual Language in Education Policy in South Africa. However, if PIRLS is not providing a valid, reliable or comparable measure of reading ability across language groups the insights that can be drawn from participation are not as meaningful.

South Africa has long recognised they have a reading crisis, the same low results have been seen through participation in other regional and national large-scale assessments, as well as through both qualitative and quantitative investigations of reading ability at national and regional levels (e.g., Ardington et al., 2020; Badat & Sayed, 2014; DBE, 2017a; Fleisch, 2008; Gustafsson, 2019; Howie, 2012, 2022; Kanjee & Sayed, 2013; Kotzé et al., 2019; Taylor, 2019; van Staden & Combrinck, 2023). This is an ongoing problem; a relic from an apartheid regime that purposefully and systemically restricted equal access to quality education on the basis of race and language. What is needed is detailed, rigorous, and accurate information about what works in reading development, where the issues lie in South Africa, and what will be effective to reduce the persisting attainment gaps. Unfortunately, the findings of this paper suggest that PIRLS Literacy 2016 was too difficult for most students in this study. This is concerning because when an instrument is unable to measure all participants effectively, the information obtained from participation is no longer valuable (Rutkowski et al., 2019).

The choice not to participate in any large-scale assessment, although sometimes recommended or preferred, is not always an option. ILSA data is often considered more robust compared to other data and in some cases funding and support from international organisations (e.g., World Bank) entails participation in certain ILSAs (Addey & Sellar, 2019; Howie, 2012, 2022; Reddy, 2010). Additionally, withdrawal from participation in a particular test (particularly after multiple cycles of poor results) may imply that the government is not committed to improving education in the country (Addey & Sellar, 2019). This is a highly relevant concern for South Africa, where the recent decline in PIRLS 2021 results has been widely discussed and elicited critique towards the lack of investment and plans from government to improve reading progress (e.g., Govender, 2023; Joint Education Trust, 2023; Maimane, 2023; Maynier, 2023; SAHRC, 2023).

Focusing on improving existing local testing programmes (e.g., national and provincial systemic assessments) in the place of PIRLS may help to alleviate some of the challenges that come with participating in ILSAs that are not well suited to the South Africa context. This would also enable the opportunity to focus more on the specific needs of the country. However, regardless of where the assessment is developed, or which models and procedures are used; an ongoing commitment to ensuring the results are valid, reliable, and comparable should be prioritised to ensure the assessment is fit for purpose.

6.7 Limitations and future research

Some of the more nuanced and complex aspects of the validity of both the data and methodology are central to this research and have been interrogated in this paper. However, there are nevertheless limiting factors that need mention. A primary limitation when dealing with secondary data such as this is that secondary researchers do not have input into the research design and framework for data collection. One of the limitations here is that due to the sensitive nature of race or ethnicity data, it is not publicly available as part of PIRLS data and therefore cannot form part of the analysis in this study. Racial segregation is a key feature in South Africa's past educational policies that continues to impact educational outcomes today so not including it in the analysis means that the impact of racial segregation cannot be fully accounted for.

Other contextual factors, such as language proficiency in the test language is also relevant. This study separated students into language groups based on the language in which they wrote the test. The national report noted that at all times attention was given to ensure that students had had exposure to the language of the test for at least four years prior to participation (Howie, et al., 2017). However, substantial research has shown that using a language as a medium of instruction does not equate to proficiency in a language. This, together with the dynamic multilingual context throughout the country, means that care needs to be taken to ensure that variance both across and within language groups is considered.

There are also limitations introduced related to the models used. This paper chose to apply IRT models that were used in the design and analysis of PIRLS internationally (i.e., 2PL, 3PL and GPCM), however, sample sizes for the three language groups meant that there was insufficient data for the stable estimation of the 3PL model. This means that the modelling used here was not entirely consistent with the original methods used by test developers and comparing the impact of the IRT model choice is a recommendation for future research. However, due to the extent of mistargeting found between South African student abilities and item difficulty levels, it is unlikely that utilising the 3PL model would have changed the overall conclusions around inadequacies in the assessment difficulty for the local context.

Future research is also recommended to address some of the issues that arose in the analysis for this paper. Firstly, this analysis considered issues of fit at an item level only. The design of PIRLS, where students only respond to two of 12 reading passages, means that investigations at a passage level would provide further insight into the overall impact of the item-level concerns found. This is particularly recommended because a number of items that had multidimensionality concerns, low response counts, and low discrimination predominantly occurred in four passages.

Finally, this paper only focused on three of the 11 language groups tested. While the Sepedi group had the smallest sample size of all 11 groups, it is important to note that the targeting issues identified are more likely related to floor effects – where the test items are too difficult for a significant portion of the students – than to sample size alone. Moreover, the consistently low scores observed in the Sepedi group were similar to overall low distributions of scores observed in the eight other African language groups, all of which had more homogeneity in performance relative to the Afrikaans and English groups. If another African language group had been selected, therefore, even with a larger sample than the Sepedi group, it is likely that similar challenges would arise due to the homogeneity of low scores observed across all nine African language groups. Future research is recommended to understand the extent to which the concerns found for the groups included in this study also apply to the eight African language groups not investigated.

CHAPTER 7 (PAPER 3)
DIF HAPPENS: THE IMPACT OF DIFFERENTIAL ITEM FUNCTIONING ACROSS LANGUAGE VERSIONS OF PIRLS 2016 IN SOUTH AFRICA

Abstract

This paper investigates the comparability of three language versions of the Progress in International Reading Literacy Study (PIRLS) 2016 in South Africa. Differential response functioning (DRF) techniques are applied to evaluate comparability at item, passage, and overall test levels for the Afrikaans, English, and Sepedi language groups. Significant differential item functioning (DIF) was found for 28% of items between Sepedi and English, and 20% between Afrikaans and English. Across the 12 PIRLS passages, 11 passages functioned differently between Sepedi and English groups, and four passages functioned differently between Afrikaans and English. While there were no significant test-level differences between the Afrikaans and English versions, significant overall divergence between Sepedi and English versions emerged. The findings suggest meaningful comparability issues between the Sepedi and English versions of PIRLS 2016 in South Africa. This has implications for interpreting reading achievement across language groups as score comparisons may be problematic. The findings highlight the need to consider within-country language comparability carefully to ensure results offer fair and valid comparisons across groups.

Key words

Differential functioning, cross-language comparability, DIF impact, PIRLS Literacy

7.1 Introduction

The Progress in International Reading Literacy Study (PIRLS) is an international large-scale assessment (ILSA) conducted in five-year cycles that evaluates reading achievement for 10-year-old students. PIRLS is designed and administered by the International Association for the Evaluation of Educational Achievement (IEA) and it aims to provide valid comparisons of reading literacy across diverse countries and languages. The 2016 administration involved over 300,000 students from 50 countries and was administered in 41 different languages. Although the comparability of different language versions of ILSAs such as PIRLS is considered carefully during the development stages, this is typically applied to the entire international sample and similar rigor is not applied to ensure the comparability of language versions within multilingual jurisdictions.

Investigating comparability in ILSAs typically involves the use of differential item functioning (DIF) analysis techniques to identify items that function differently for students from different groups with the same overall ability level (e.g., Ercikan & Por, 2020; Goodrich & Ercikan, 2019; Mtsatse & van Staden, 2021). However, the impact of DIF at the item level can be difficult to interpret at the overall test level (Chalmers, 2022; Meade, 2010). Although the individual effect of items flagged for DIF may be small, cumulative effects can substantially advantage or disadvantage groups (Chalmers et al., 2016). Therefore, researchers tend to consider the extent to which DIF impacts the overall test scores for a particular group, and what might lead to DIF.

PIRLS 2016 was conducted in 11 language versions in South Africa. Its comparability across language versions is an important consideration because the overall reading performance of students across these language groups differed considerably (Howie, Combrinck, Roux, Mokoena, et al., 2017). Students taking the test in an African language severely underperformed compared to those tested in English or Afrikaans (Howie et al., 2017). The disparities seen in PIRLS results reflect the inequality still present in South Africa's education system stemming from decades of racial and linguistic discrimination (Spaull & Pretorius, 2019; van Staden & Combrinck, 2023). Added to this, the average overall performance of South African students across all language groups was substantially lower than the international average, which has the potential to threaten the reliability and validity of the results (Howie, 2012; Lockheed & Wagemaker, 2013; Rutkowski et al., 2019). Given the potential policy implications of results from large-scale assessments such as PIRLS, it is essential to rigorously evaluate whether the assessment provides truly comparable and fair measures across linguistically diverse student populations such as South Africa before making decisions based on the scores.

The diverse, heterogenous linguistic landscape of South Africa means that any discussion of reading achievement needs to consider the multilingual context carefully (DBE, 2023). Ensuring within-country comparability across language groups is therefore critical to better understand reading achievement in South Africa. This paper presents an analysis of PIRLS 2016 data from South Africa, focusing on differential functioning at item, passage, and overall test levels across 3 of the 11 language groups.

7.1.1 Aim and research questions

The primary aim of this study is to determine the extent of comparability across three language versions of PIRLS Literacy 2016 in South Africa. To examine this, the study applies Chalmers' (2018) differential response functioning (DRF) methodology to quantify the accumulated effect of item-level DIF and determine if significant overall bias exists between language versions of PIRLS 2016 in South Africa at both a passage and overall test level.

The Afrikaans, English, and Sepedi versions have been selected for analysis due to the large differences in scores across the highest (Afrikaans or English) and lowest (Sepedi) performing groups. Moreover, the existing research in into DIF in South Africa has focused primarily on more commonly spoken Nguni languages like isiZulu and isiXhosa. There is limited research on reading in African languages broadly, and a particular gap around Sotho languages. This is concerning given the two lowest performing groups in PIRLS 2016 in South Africa were Sotho languages; Sepedi and Setswana. Given this performance gap, research into the assessment of languages from the Sotho family is particularly important.

To understand the extent of potential comparability issues, three research questions are investigated:

1. To what extent are the Afrikaans, English, and Sepedi versions of PIRLS Literacy 2016 comparable at an item level?

2. To what extent does the comparability of PIRLS Literacy 2016 across these languages at an item level impact on the comparability of the assessment at a passage level?
3. To what extent does the comparability of PIRLS Literacy 2016 across these languages at item and passage level impact on the comparability of the assessment at test level?

7.2 Literature review

7.2.1 Ensuring quality in international large-scale assessments

ILSAs such as PIRLS are designed to provide comparisons of student achievement across diverse countries and language groups. Guidelines from the International Test Commission (ITC, 2001, 2017, 2019) and Standards for Educational and Psychological Testing (AERA et al., 2014) outline comprehensive frameworks and sets of criteria to ensure quality in large-scale assessments. Central to these guidelines is the notion that tests need to ensure that the proposed construct is being validly, reliably, and fairly measured across all intended contexts and for all intended interpretations and uses (Bartram & Hambleton, 2016).

One of the fundamental principles of quality assessment is validity. Although the conceptualisations and definitions of validity have changed over the years; it is broadly acknowledged to refer to the degree to which both evidence and theory support the interpretations of test scores for its proposed uses (Kane, 2013; Messick, 1998; Newton, 2012; Newton & Shaw, 2014). Added to this, issues of fairness and reliability are fundamentally intertwined with validity and when either reliability or fairness is compromised then validity is also compromised (Newton & Shaw, 2014). Ensuring the validity of a large-scale assessment like PIRLS, therefore, needs to account for the intended purpose of the test in the various contexts in which it is used, the reliability of the measurement across groups, and the fairness of the test across groups with diverse characteristics (Embretson, 2007; Hambleton et al., 2004; Sireci, 2007).

Establishing robust validity evidence for ILSAs like PIRLS is challenging given the multifaceted constructs measured and diverse contexts involved. To substantiate validity, PIRLS implements extensive quality assurance checks to detect unusual item properties that could indicate translation, printing, or scoring issues for particular countries (Foy et al., 2018). Added to this, they employ conditioning procedures that incorporate background data to ensure that the proficiency estimates are as reliable, accurate, and unbiased as possible while maintaining underlying grouping structures in the data (Foy & Yin, 2017). PIRLS also employs rigorous translation and adaptation methodologies that aim to ensure the instrument reflects the most accurate and comparable picture of student performance possible across versions (Hambleton & Lee, 2013; Martin, Mullis, & Hooper, 2017; Nilsen et al., 2022).

Nevertheless, researchers continue to raise doubts about the extent to which ILSAs are able to meet these objectives. A number of researchers have questioned whether the differences in ability ranges across groups affect validity, reliability, and fairness of the instruments across countries (e.g., Atar et al., 2023; Rutkowski & Rutkowski, 2021; Wagemaker, 2020). There are also concerns raised about the extent to which procedural, analytical, or design choices affect the comparability of the instrument across groups (Ercikan & Por, 2020; Kreiner & Christensen, 2014). A key challenge raised regarding ILSAs surrounds the conceptual and practical feasibility of measuring complex constructs fairly and validly across different languages and cultures (e.g., El Masri et al., 2016; Ercikan & Koh, 2005; He et al., 2018; McGrane & Maul, 2020).

7.2.2 Cross-language comparability

Exploring the comparability of difficulty across different versions of large-scale assessments has been one of the most active areas of psychometric research for decades (Jones & Thissen, 2006). It is widely held that the factors that affect test comparability go beyond just an evaluation of form and properties; ensuring that language versions represent the same level of difficulty for each group is fundamental to fair and valid comparisons (Arffman, 2010; Hambleton et al., 2004). Added to this, the impact of language-specific differences is increasingly being recognised as an important consideration for cross-language assessment comparability (El Masri et al., 2016; Ercikan & Por, 2020; Gökçe et al., 2021; Hambleton & Lee, 2013; Mtsatse & van Staden, 2021).

Investigations into comparability concerns that focus specifically on PIRLS have revealed evidence that different versions do not function comparably for some language groups (e.g., Goodrich & Ercikan, 2019; Grisay et al., 2009; Mtsatse & van Staden, 2021; Roux et al., 2022). One of the most widely used approaches to evaluate the comparability of different language versions of ILSAs is DIF techniques (Li et al., 2022). These are used to identify where students from different groups with the same underlying ability on a particular test, have a different probability of answering an item correctly (Osterlind & Everson, 2009; Zumbo, 2007). The presence of DIF, therefore, threatens test validity by indicating items are not measuring constructs consistently across groups. For example, a DIF analysis by Sandilands et al. (2013) investigating the comparability of English and Spanish versions of PIRLS 2001 across three countries found that between 21% and 38% of items functioned differently between language groups. In Canada, Goodrich and Ercikan (2019) evaluated the comparability of French and English versions of PIRLS 2011 and found that up to 25% of items showed evidence of different levels of difficulty across language versions. Meanwhile, Grisay et al. (2009) found that across all national versions of PIRLS and the Programme for International Student Assessment (PISA), 20% of the variance in item difficulty was attributable to item-by-language interaction.

A further challenge is ensuring comparability between different language versions within a country, particularly in a context like South Africa where performance across language groups differs substantially. Between 43-44% of students who took PIRLS 2016 in Afrikaans or English achieved the basic reading level in PIRLS, whereas only 7-18% of students who took the test in one of the nine African languages were able to achieve the same level (Howie, Combrinck, Roux, Mokoena, et al., 2017). Researchers looking at the comparability of PIRLS within South Africa have found evidence of items that function differently for students across language versions. In an investigation comparing the English and isiXhosa versions of a passage from PIRLS 2011, Mtsatse and van Staden (2021) found evidence of DIF, with 3 of the 15 items in the passage functioning differently for students across language groups. Roux, van Staden and Pretorius (2022) investigated issues of equivalence across Afrikaans, English, and isiZulu versions of three passages for PIRLS 2016, revealing that 20-38% of items in PIRLS Literacy 2016 functioned differently across the three language versions. Combrinck (2020) looked at issues of measurement invariance for the isiZulu versions of PIRLS in both 2006 and 2016 and found that 18-23% of the items that were common to both cycles functioned differently across the two cycles.

In addition to identifying DIF items in a test, understanding the sources or causes of DIF is important for informing item and test development. This typically involves explanatory analysis where language experts, translators, and content specialists review items to identify potential sources of DIF based on features, content, topics, or cultural references that could disadvantage a group (Allalouf et al., 1999; Finch & Hernández Finch, 2016; Li et al., 2022). Several researchers have adopted this approach in an effort to uncover the underlying sources of DIF in ILSAs (e.g., Allalouf et al., 1999; El Masri & Andrich, 2020; Ercikan et al., 2010; Finch & Hernández Finch, 2016; Huang et al., 2016; Li et al., 2022; Oliveri et al., 2013; Sandilands et al., 2013). One of the challenges with exploratory approaches to understanding DIF is that a number of studies have found that although expert judgement is able to provide insight into potential linguistic or contextual inconsistencies across items; the review procedures are not always effective in predicting DIF items or isolating their specific causes (Gorsky et al., 2019; Li et al., 2020; Roth et al., 2013). Added to this, while informative for future test improvements, explanatory analysis has limitations in evaluating the cumulative impact of DIF found on overall score comparability.

7.2.3 The overall impact of differential item functioning

Another approach to examining DIF, therefore, involves evaluating the overall impact of DIF identified across groups. This evaluative approach, referred to as differential test functioning (DTF), involves quantifying the cumulative impact of DIF on score comparability across groups (Wainer, 1993). Numerous studies have shown that while individual items may display statistically significant DIF, the effect on total test scores can be negligible if DIF is small or balanced in both directions (e.g., El Masri & Andrich, 2020; Ercikan & Por, 2020). However, although instances of DIF may be small at the item level, their aggregated effect across the test can result in meaningful disadvantage for a particular group (Chalmers, 2022; DeMars, 2011; Heinz et al., 2022). DTF analysis, therefore, is an important consideration as it can determine the overall impact of small or large DIF in terms of meaningful score differences across groups (Meade & Wright, 2012; Steinberg & Thissen, 2006).

The evaluative approach to understanding DIF begins with detecting significant DIF using common techniques such as the Mantel-Haenszel methods, logistic regression or item response theory (IRT) approaches (Hunter, 2014; Li et al., 2022; Teresi et al., 2021). Effect sizes for the identified DIF are then quantified to determine the magnitude of the DIF where it exists. DTF procedures then include statistical hypothesis tests and effect size measures to assess the significance of overall differential functioning between groups (Chalmers, 2018; Kleinman & Teresi, 2016; Meade, 2010; Raju et al., 1995; Wright & Oshima, 2015).

Recent advancements by Chalmers et al. (2016) and Chalmers (2018) have strengthened DTF methodology through the use of statistics calibrated to IRT frameworks which enable the statistical tests to better handle uncertainty and more accurately quantify the impact of DIF at a test-level (Bauer, 2023; Schneider et al., 2022; Temel, 2023; Teresi et al., 2021). By integrating DIF and DTF techniques, the evaluative DTF approach provides essential validity evidence regarding score comparability that can guide result interpretations.

7.3 Methods

7.3.1 Data

The data for this study was retrieved from the PIRLS 2016 international database and is a subset of 4,215 students from the national sample (N = 12,810) of South Africa. Student responses were included in the analysis if students took PIRLS Literacy 2016 in Afrikaans (n=1,228), English (n=2,089) or Sepedi (n=898). Based on the mean PIRLS literacy scores, the English group scored highest 371 (SD 116), followed by Afrikaans 351 (SD 118), and Sepedi 274 (SD 87). The variability in each group differed, with the Sepedi group showing the narrowest spread of scores, and the English and Afrikaans groups both having similar, more widely spread, distributions. The international PIRLS average is standardised to have a mean of 500 and standard deviation of 100. Therefore, the average of all three groups was at least one standard deviation below the international average. Figure 7-1 shows the distribution of PIRLS achievement scores for the three groups, averages for each group are indicated by the dashed lines, and the international mean is shown in red.

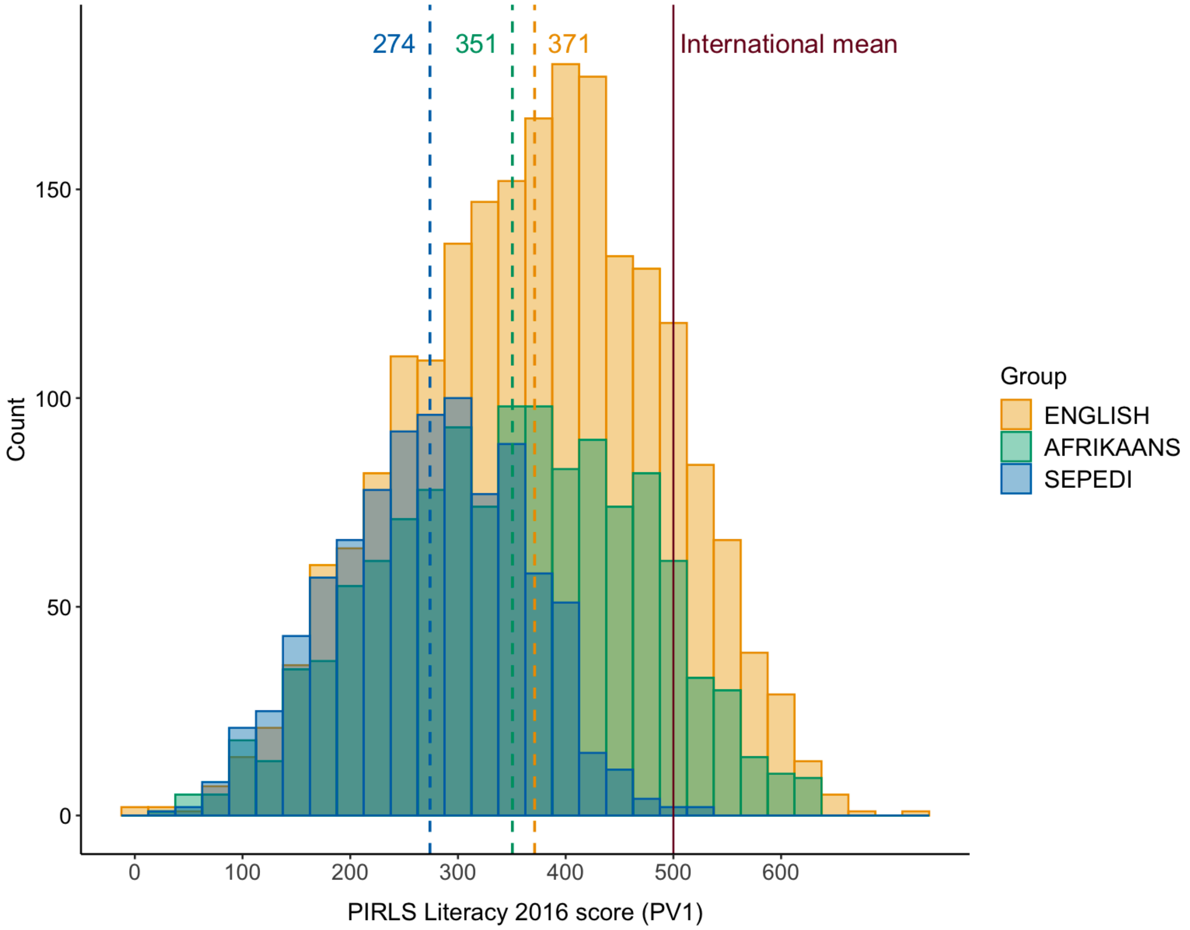


Figure 7-1 Histogram of average PIRLS Literacy 2016 scores for each language group

7.3.2 Passages and items in PIRLS Literacy 2016

The passages that form part of PIRLS are separated into two categories: literary texts, and informational texts. The literary passages include contemporary short stories, traditional tales, and fables, while the informational texts draw from various sources, resembling textbook or article formats with text and graphical information. The selection of passages was subject to qualitative evaluation by experts at the IEA and National Research Coordinators (NRCs) from participating countries, assessing factors like text length, readability, and cultural adaptability for different languages (Martin, Mullis, & Hooper, 2017). Additionally, quantitative evaluation metrics, including word count and the Flesch-Kincaid Grade Level Formula, were provided to NRCs to gauge text readability. Passages are between 400 and 800 words and there are about 15 accompanying items per passage, interspersed within the text to assist students in retrieving information as they read.

To manage the assessment workload, students are required to respond to items related to just two passages, distributed across booklets with an expected testing time of 40 minutes. The design of the booklets is aimed at ensuring a fair distribution of passages among students and they are administered using a matrix block design (Martin, Mullis, & Foy, 2017). Ten blocks were created, five for literary and five for informational content, and these were systematically allocated to 15 booklets. This approach ensures that each passage appears three times but in different combinations.

The students in this dataset participated in PIRLS Literacy 2016 which is designed to be an easier version of the main PIRLS 2016 test that is still scalable with the main test. One strategy employed to ensure comparability between the two versions is the integration of passages from the main test into the PIRLS Literacy booklets, referred to as "linking" passages. As passages in PIRLS are more difficult than those in PIRLS Literacy, when passages shared with PIRLS are included in the booklets, they always appear in the second position. The 16^{th} booklet, referred to as the Reader, is designed in the format of a magazine to present a more authentic reading experience for students. It contains a literary and informational passage that do not overlap with the other booklets but are shared across PIRLS and PIRLS Literacy 2016. To maintain proportional distribution, the Reader booklet is assigned a higher frequency during random distribution, ensuring an equal response rate compared to other booklets.

The item set from PIRLS Literacy 2016 includes 183 items across these 12 passages; 91 multiple choice items and 92 constructed response items; 26 polytomous (scored 0,1,2 or 0,1,2,3), and 157 dichotomous (0,1). Table 7-1 shows the passages including the number of items and types of items in each passage.

Table 7-1

Summary of passages and items in PIRLS Literacy 2016

Passage	Passage title	Passage type	Number of items
A	Ants	Informational	14
B	Bhagita's perfect orange	Literary	16
C	Training a deaf polar bear	Informational	14
D	The summer my father was ten	Literary	14
E	Pemba sherpa†	Literary	17
F	African rhinos & oxpecker birds	Informational	17
G	How did we learn to fly?†	Informational	17
H	Hungry plant	Informational	16
I	The pearl	Literary	15
J	Library mouse	Literary	18
K	Flowers on the roof†	Literary	13
L	Sharks†	Informational	12

Note: † indicates passages shared with main PIRLS test

7.3.3 Detecting differential item functioning

An IRT framework was used as the approach for estimating student and item parameters in this study which could then be compared to determine if they function differently across language groups. Similar to the approach used by PIRLS 2016[10], a two-parameter logistic model (2PL model) was used to estimate student and item parameters for dichotomous items, and a Generalised Partial Credit Model (GPCM) was used for polytomous items. An IRT-based method for identifying DIF using likelihood-ratio tests, IRT-LR (Thissen, 2001), was selected as it is able to provide both common and

[10] Although the 3PL model was also used for PIRLS 2016, specifically for estimating item parameters for the multiple-choice items, the reduced dataset in this study does not provide sufficient sample size for accurate and reliable estimation of guessing parameters (Embretson & Reise, 2013). As a result, the 2PL model was used to estimate parameters for all dichotomous items including multiple-choice items.

group-level item parameter estimates, identify both uniform and nonuniform DIF, and can be applied to dichotomous and polytomous item types (Tay et al., 2014). Added to this, in assessments where the construct is complex, IRT-LR has been shown to be a balanced and effective method of detecting DIF (Bulut & Suh, 2017). One of the challenges when using likelihood-ratio tests for DIF detection is that they are prone to being affected by large sample sizes. Although attempts to moderate these effects through Benjamini-Hochberg (B-H) adjustment (1995) as well as the anchor linking technique are applied during DIF detection, it may be the case that the magnitude of DIF identified may be negligible in terms of the overall effect size (Haynes, 2013; Meade & Wright, 2012; Wang & Woods, 2016).

The magnitude of the effect size of DIF for each DIF item was therefore calculated using the approach suggested by Meade (2010), who recommends considering multiple measures of effect size to understand any potential DIF effects fully. For this study, three measures of item effect size were considered: signed item differences in sample (SIDS), unsigned item differences in sample (UIDS), and expected score standardised difference (ESSD) (Meade, 2010; Tay et al., 2014). ESSD was selected as the main index to classify DIF effect size as it provides a standardised effect size in the metric of the overall expected score on each item, this means that meaningful interpretations of the potential impact of DIF on students in the focal group can be made in the same unit as the scoring for each item.

As with ESSD, SIDS also provides an index on the metric of the expected score for each item, and by comparing it with UIDS further information about the nature and direction of DIF can be uncovered. Where SIDS and UIDS differ, DIF is affecting students across groups differently according to where they fall along the ability scale (i.e., non-uniform DIF). Uniform DIF, on the other hand, occurs when one group of students finds an item more difficult than the other group across the distribution of student ability.

7.3.4 Evaluating the impact of differential item functioning

Once the presence, magnitude and direction of DIF was established at an item level, the overall impact of DIF was evaluated at both the passage and test level. DIF at a passage and test level may occur because a large number of items in a test contain DIF for a specific group, or in situations where there are not necessarily items that show large instances of DIF independently but, when considered cumulatively, small differences in item functioning in the same direction for one group may aggregate and result in the test functioning differently for that group (Chalmers, 2018; Chalmers et al., 2016; Temel, 2023). Considering the impact of DIF at a passage level is a particularly important concern for PIRLS as the assessment design means that each student only responds to a selection of passages that make up the PIRLS reading test. As a result, if systematic differences in difficulty are found at a passage level across groups, this could impact the overall comparability of the test because some groups of students may have responded to passages that presented a systematically higher or lower level of difficulty than others.

To understand the impact of differences in difficulty across passages and overall on the test, the DRF statistics proposed by Chalmers (2018) are used. A benefit of using these statistics to understand comparability at the passage and test level is that they do not rely only on the overall test score to determine student ability, instead comparisons can be made that account for item parameter estimates at varied ability estimates (Chalmers, 2022). Furthermore, relative to other test-level approaches such as the compensatory SIBTEST approach (Shealy & Stout, 1993) and the DFIT framework (Oshima et al., 1998; Raju et al., 1995), DRF statistics are not as severely impacted by Type 1 error rates and they are more effective when groups differ in ability (Finch & French, 2023; Temel, 2023; Teresi et al., 2021). This is particularly important for this study as there is substantial variation in student ability between groups. In addition, DRF statistics have been shown to present a more robust and reliable evaluation of the overall impact of DIF in situations where DIF is non-uniform (i.e., in situations where DIF disadvantages groups differently at different ability levels) (see Bauer, 2023; Chalmers, 2022; Temel, 2023).

In this study the impact of DIF was quantified using two statistics from the DRF framework; the signed differential response function (sDRF) and the unsigned differential response function (uDRF) (Chalmers, 2018). sDRF refers to DIF that favours one group, in other words when items function differently in one direction, while uDRF represents the absolute values of the differences between groups irrespective of the direction of the difference. sDRF represents the expected score differences (S) between the item parameters (ψ) of the focal (F) and reference (R) groups, averaged across the ability distribution $f(\theta)$. uDRF sums the absolute score differences between the focal and reference groups measuring their average deviation regardless of the direction of the difference. uDRF values closer to 0 signify more similar expected performance between groups, while higher uDRF values indicate greater dissimilarity in expected group performance. The equations that follow represent each of the DRF statistics.

Equation 3: Signed differential response function

$$sDRF = \int [S\left(C\middle|\psi^{(R)}, \theta\right) \mathrm{S}\left(C\middle|\psi^{(F)}, \theta\right)] f(\theta)\, d\theta$$

Equation 4: Unsigned differential response function

$$uDRF = \int |S\left(C\middle|\psi^{(R)}, \theta\right) \mathrm{S}\left(C\middle|\psi^{(F)}, \theta\right)] f(\theta)\, d\theta \mid f(\theta) \mathrm{d}\theta$$

In both equations, $f(\theta)$ is empirically estimated from the posterior distribution using an empirical histogram (EH) approach rather than relying solely on a normal distribution (Chalmers, 2018). This allows for non-normality in the data and prevents overreliance on assumptions about the ability distributions that may not match the data (Chalmers, 2022). In addition to statistical comparisons of passage functioning using DRF statistics, the magnitude of differential functioning found at both a passage and test level were described using the values of sDRF and uDRF.

A benefit of these statistics is that they are in the same metric as the expected scores for each passage. This means that they are straightforward to interpret in terms of actual impact of DIF identified for the students in each group. For instance, if the maximum score on a passage is 10 points, an sDRF value of +.5 means the focal group has a 5% higher probability of obtaining each specific score on that passage compared to the reference group, while a uDRF value of 1.5 indicates a 15% absolute difference in probability between groups. This enables clear interpretation of the effect of differential functioning across groups at an overall test level by providing a measure of the total score difference between groups with or without DIF present (Gnambs & Lenhard, 2023; Schöfl et al., 2023).

7.4 Analysis

All IRT analyses were conducted using the mirt package version 1.41 in R which applies the marginal maximum likelihood method with an expectation-maximum algorithm to estimate models (Chalmers, 2012). As recommended by Chalmers (2018), the fit of the data to the IRT model was evaluated using comparisons of Akaike Information Criterion (AIC) and the Bayesian Information Criteria (BIC) (which accounts for additional parameters in the model), as well as a likelihood ratio test (χ^2) to compare two unidimensional models; a complex IRT model (2PL model and GPCM) and a simple IRT model (Rasch model). By comparing a set of competing models for the same data – the best fitting model can be identified (Burnham & Anderson, 2002).

7.4.1 Item fit evaluation

Given South Africa's low performance in PIRLS 2016 compared to the international sample, there was concern that some test items may be too difficult for this subset of students. When test items are too difficult, this can lead to issues with model fit and targeting that threaten the reliability and validity of the test (Rutkowski & Rutkowski, 2021; Zhao & Hambleton, 2017). Furthermore, items that are poorly targeted with students' ability also threaten the accuracy and precision of DIF detection when using IRT-based methods (DeMars & Jurich, 2015). Therefore, prior to conducting the DIF analysis, a thorough analysis of the full set of PIRLS 2016 items was conducted to evaluate whether there were items that were misfitting for this subset of students.

The item fit evaluation considered the proportion of correct responses, violations of local independence, discrimination values, and a graphical inspection of the fit between the expected and theoretical item response functions (IRF). Items were flagged for misfit if they had fewer than 15 correct responses per language group, violated the assumption of local independence according to Yen's (1984) Q_3 statistics (at $Q_3 > .20$), or violated the assumption of unidimensionality by showing no relationship with the primary factor in the Item Factor Analysis (at F1 < .32) (Tabachnick et al., 2007). Additionally, items were flagged for misfit if they had low discrimination values estimated by the IRT model ($a < .25$), or if the observed responses did not correspond with the expected fit of the IRT model through graphical inspection of IRFs.

7.4.2 Differential item functioning analysis

In order to test for differential functioning at an item level, three multiple group models were estimated to allow for significance testing using likelihood ratio tests, a constrained model (M_0), a free model (M_1) and an anchored model (M_2). The results of comparisons between M_0 and M_1 revealed 87 items eligible for inclusion in the anchor set and from this, as suggested by Meade and Wright (2012), 15 items (~10% of total items) with the highest discrimination parameters, and with no evidence of previous fit issues, were selected to form the anchor set (for comparisons with M_2).

A second round of IRT-LR tests were then conducted pairwise using the anchored model with English as a reference group and Afrikaans and Sepedi as focal groups. Items were flagged for DIF if the B-H adjusted p-values in the second round of testing showed significant differences in the item difficulty between reference and focal groups at a 99% confidence level ($p < .01$). Once DIF items had been identified for each focal group, the magnitude of the differences was quantified in terms of three indices of effect size: ESSD, SIDS and USIDS (Meade, 2010; Meade & Wright, 2012). ESSD is a standardised measure of effect size that uses the metric of the expected score for each item. DIF items for each focal group have been classified according to four DIF effect sizes according to their ESSD index, negligible DIF (ESSD < .2), small DIF (.2 < ESSD < .5), medium DIF (.5 < ESSD < .8) and large DIF (ESSD >. 8).

7.4.3 Differential item functioning impact analysis

Following the item level analysis, the impact of DIF found was evaluated at both the passage and test levels. The passage level comparisons were conducted by 'bundling' together all items associated with each passage and evaluating them as a set using Chalmers' (2018) framework of DRF statistics. Test-level DRF analysis was undertaken by evaluating all items in the PIRLS Literacy 2016 assessment. In both cases, comparisons were conducted pairwise with English as the reference group. The item parameters from the previously estimated anchored multiple group model (M_2) were used to estimate sDRF and uDRF statistics for each comparison pair. To account for uncertainty in the estimated item parameters, the DRF analysis was repeated (1000 replications) using different sets of plausible item parameters for each passage and for the test (see Chalmers, 2018). These were randomly generated based on the amount of uncertainty in the original estimates. This approach, called parametric bootstrapping, provides a robust evaluation of differential passage and test functioning that acknowledges the inherent uncertainty in the estimated IRT parameters. From the parametric bootstrap results, 95% confidence intervals were empirically constructed around the mean DRF statistics.

DRF statistics were compared graphically and differences across groups were flagged as showing significant differential functioning if the signed (sDRF) or unsigned (uDRF) p-values were below .01 based on likelihood-ratio testing. A significant DRF statistic indicates that the effects of DIF found in the items in a passage or test result in a systematic difference in the level of difficulty presented by that passage or test across the groups.

7.5 Results

7.5.1 Item response theory model estimation and fit evaluation

The model fit evaluation identified the complex IRT model (2PL model and GPCM) as the best fitting model (χ^2 (182) = 2968, $p < .001$) (see Table 7-2). The complex IRT model was therefore selected to estimate item parameters for all groups so that any differences in item functioning – in terms of their difficulty parameters – could be evaluated across groups. Prior to conducting the DIF analysis, IRT model assumptions and item fit were scrutinised carefully to account for issues related to targeting or potential flooring effects.

This evaluation of item fit found that 31 items showed evidence of misfit or poor targeting for the groups of students in this study. Specifically, 14 items had fewer than 15 correct responses in any language group and were removed as these can severely impact the accuracy of model estimation. An additional 17 items were removed as they showed evidence of misfit in two or more of the following checks; they violated the assumptions of the model, showed poor overall fit with the model expectations, or they had very low discrimination values. Removing these 31 items provided a reduced set of better functioning items to enable more reliable DIF analysis between groups. While reducing items risks compromising construct coverage, retaining items with evidence of misfit would threaten the reliability and accuracy of the DIF results. More details regarding the fit checks undertaken and decisions about the removal of items is covered in Section 6.5 in Paper 2.

The reduced item set comprised 78 multiple choice items and 68 constructed response items; 144 of which are scored dichotomously and 8 polytomously (0,1,2). Following the reduction of the item set, model fit was re-evaluated and the complex IRT model remained the best fitting model at a $p < .01$ level (χ^2 (151) = 1910, $p < .001$), and had lower AIC and BIC values than those of the simple model. The reduced item set of 152 items was therefore used for all further analyses. Table 7-2 summarises the results of the model fit evaluation before and after reduction of the item set.

Table 7-2

Model fit statistics before and after item set reduction

Model	AIC	BIC	χ^2	df	p-value
Full item set (n=183)					
Simple	145498	146849	-	-	-
Complex	142894	145401	2968	182	< .001
Reduced item set (n=152)					
Simple	118287	119308	-	-	-
Complex	116687	118667	1901	151	< .001

7.5.2 *Determining comparability at an item level*

Following evaluation of the fit of the complex IRT model, the item difficulty estimates from each group were compared to identify DIF items. DIF detection for the first focal group, Afrikaans, revealed 30 items where the items functioned differently for the Afrikaans group compared to the English group at the $p < .01$ level. This means that about 20% of items presented a different level of difficulty for students taking the test in Afrikaans compared to students of the same overall ability level but taking the test in English.

When Sepedi was used as the focal group, the likelihood-ratio tests revealed a higher number of items that functioned differently at the $p < .01$ level. In total, 42 items showed evidence of functioning differently for students taking the test in Sepedi compared to those taking the test in English. This means that 28% of items presented a different level of difficulty for students from the Sepedi group with the same overall ability level as those in the English group. Of the 30 DIF items for the Afrikaans group, 18 were small/negligible in terms of effect size, while 10 items showed medium DIF, and 2 items had large DIF. There is evidence of larger effects for the DIF items for the Sepedi group. From the 42 DIF items, 18 had small/negligible effect sizes, while 10 showed medium DIF and 14 items had large DIF.

These effect sizes do not describe the direction of advantage or disadvantage for the focal group, they only describe the magnitude of the difference between the level of difficulty for each group. To understand the direction of the disadvantage, DIF items were then further described in terms of the nature of DIF (i.e., uniform or non-uniform), which was determined based on differences between the UIDS value and the absolute value of SIDS. The item-level analysis identified a high number of items exhibiting non-uniform DIF for both the Sepedi (32 of 42 items) and Afrikaans (20 of 30 items) focal groups compared to English. Non-uniform DIF indicates that the bias does not consistently advantage one group across ability levels, but rather advantages certain groups for lower proficiency students while favouring the other group among higher proficiency students. Table 7-3 shows the proportion of non-uniform and uniform DIF items found for each focal group, as well as the variation of effect size for each DIF type.

Table 7-3

Summary of effect size and direction of differential item functioning by focal group

	Afrikaans (focal)		Sepedi (focal)	
Effect size	**Non-uniform**	**Uniform**	**Non-uniform**	**Uniform**
Small/negligible	13	5	18	0
Medium	6	4	9	1
Large	1	1	5	9
Total	20	10	32	10

Note: Reference group is English in both cases

As evident in the table, a higher number of DIF items for both focal groups were non-uniform. The high incidence of non-uniformity has important implications for understanding DIF impact. Specifically, it suggests there may be "cancelling out" of bias across ability levels that could mask disadvantage at the test level if only directional DIF is considered. This reinforces the importance of evaluating both directional and absolute DIF impact on the passages and test level scores to fully examine comparability concerns (Temel, 2023).

Figure 7-2 shows the group IRFs for two items that showed large DIF when Afrikaans was the focal group (Item136 and Item87). Item136 (left) shows an example of an item exhibiting uniform DIF favouring the English group across all ability levels. Item87 (right) shows an example of a non-uniform DIF item where the IRFs cross. In the non-uniform example, the Afrikaans group is favoured at lower ability levels but as the ability of the students increases, so the direction of advantage shifts in favour of the English group.

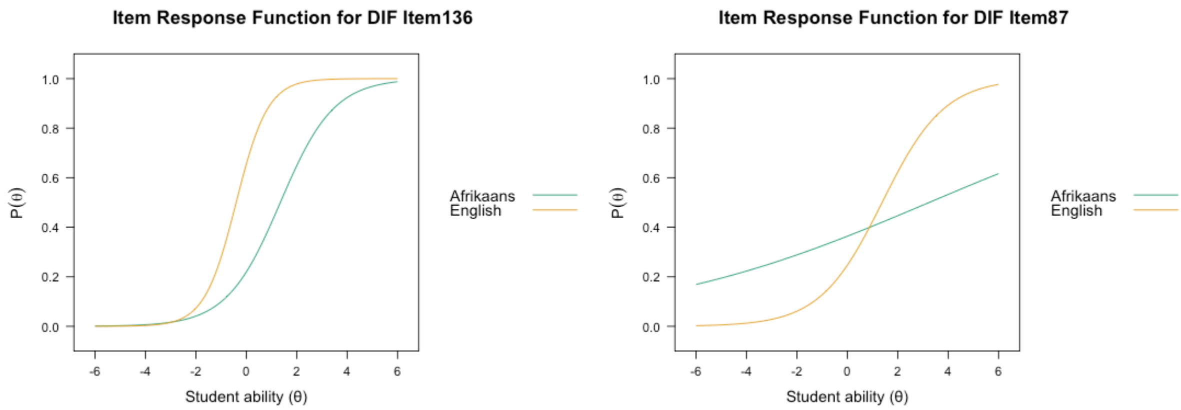


Figure 7-2 Items with large differential item functioning – Afrikaans focal group

In Figure 7-3, two selected examples from the 14 items with large DIF for the Sepedi group are shown (Item60 and Item50). Item60 (left) graph depicts uniform DIF advantaging the Sepedi group across the ability range, while Item50 (right) shows non-uniform DIF with the advantage shifting to favour the English group as ability increases. This pattern of advantage for lower-ability students in the Sepedi group was almost exclusively seen in multiple-choice items, which suggests that there may be differences in guessing behaviour across groups.

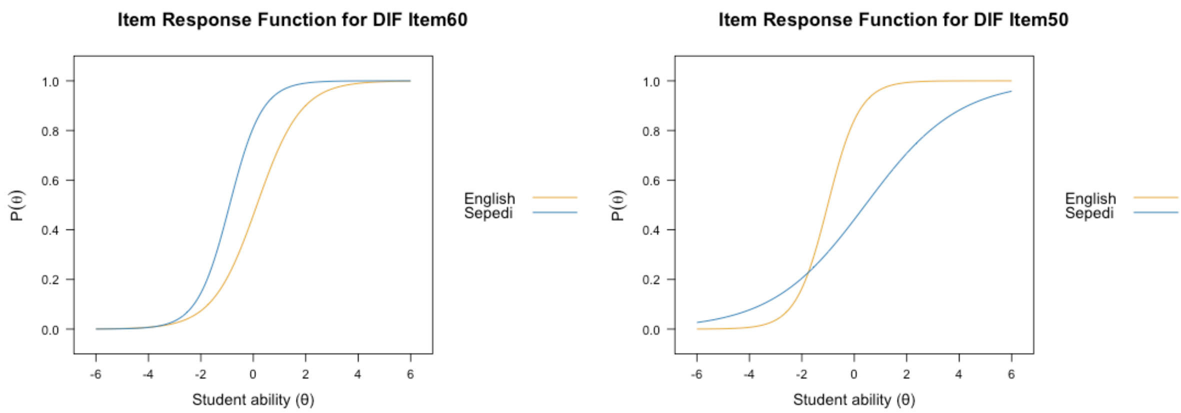


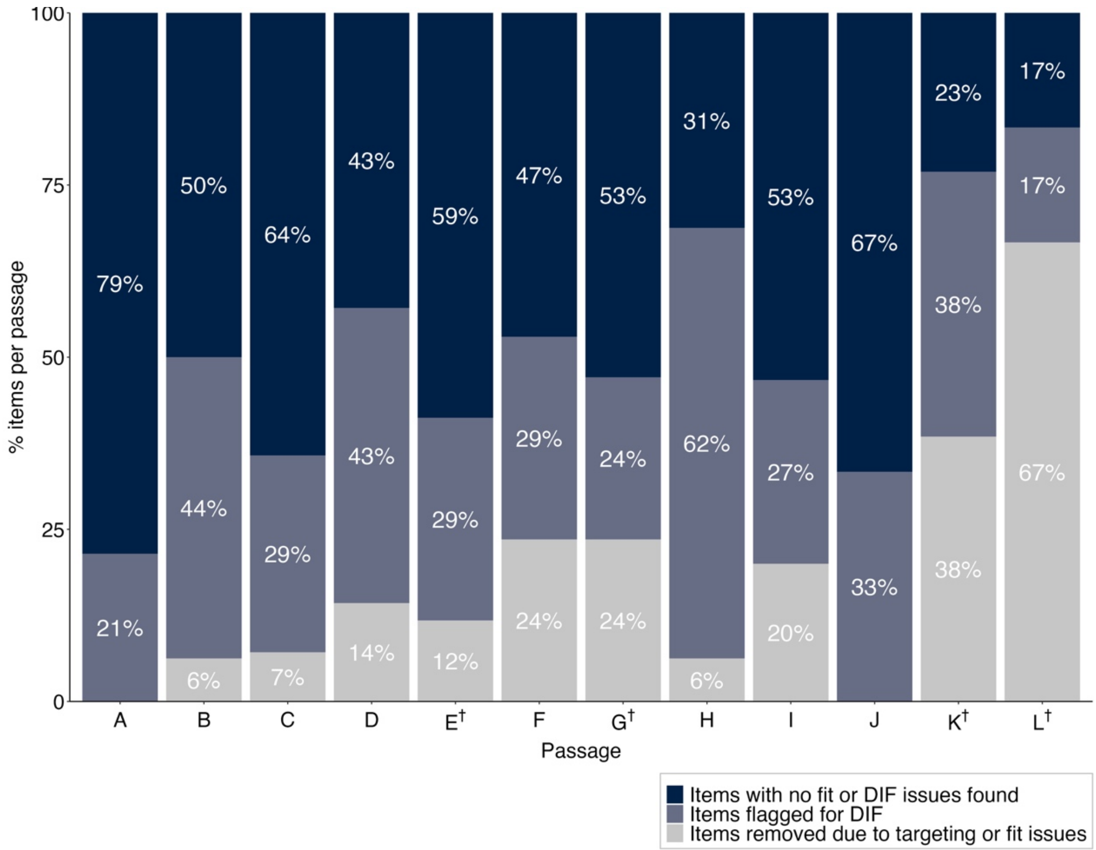
Figure 7-3 Examples of items with large differential item functioning – Sepedi focal group

7.5.3 Evaluating differential item functioning impact at a passage level

Before considering the results of the passage level analysis, it is important to note that the reduction in itemset which resulted from the fit checks affected passages differently. Figure 7-4 presents a summary of the percentage of items in each passage that were removed due to concerns with fit or targeting (light grey) and the percentage of items that had DIF flagged at an item level for either the Sepedi or Afrikaans focal group (dark grey). Also shown in this figure is the proportion of items in each passage that did not have fit issues or DIF items (dark blue), i.e., items that functioned comparably across the groups. A summary of the number of items in each passage, as well as the number of items with DIF and fit issues found per passage is included in the table below the figure.

The figure shows that there was variability in item removal rates across passages. The two passages most severely affected by the fit checks were Passage K and Passage L, with 38% (5 of 13) and 67% (8 of 12) of items excluded respectively. Although there are four linking passages, these two passages (K & L) are passages from the main PIRLS test intended to link to the easier PIRLS Literacy version and, therefore, have a higher level of difficulty by design. The other two shared passages (E & G) are easier PIRLS literacy passages included in the main PIRLS test. A majority of items that failed fit checks in Passages K and L had low numbers of correct responses (< 15 per language group) and/or targeting issues. A possible interpretation for the dominance of fit issues in these passages specifically is that the (more difficult) linking passages from the main PIRLS test are poorly targeted to students in this dataset. The disproportionate number of misfitting items raises concern about the extent to which their linkage function may be undermined. These concerns are further evidenced when considering the number of DIF items in these two passages. In Passage K only 3 of the 13 items, and in Passage L just 2 of 12 items functioned comparably with no evidence of fit, DIF or targeting issues. However, the high number of items excluded in these passages also means that the overall impact of DIF at a passage level should be interpreted with caution as a large proportion of items in the passages were not included in the DIF impact analysis.

The proportion of DIF flagged in each passaged varied from 17% in Passage L (2 of 12 items) to 62% in Passage H (10 of 16 items). Passage A had no items removed due to misfit and one of the smallest proportions of DIF items (21%). Similarly, there were no issues with fit or targeting seen in the items in Passage J, however, a third of items were flagged for DIF (6 of 18 items). To better understand the impact of the disproportionate number of DIF items per passage, the DRF statistics for each passage were computed.



	A	B	C	D	E†	F	G†	H	I	J	K†	L†
Total items	14	16	14	14	17	17	17	16	15	18	13	12
Fit issues	0	1	1	2	2	4	4	1	3	0	5	8
DIF items	3	7	4	6	5	5	4	10	4	6	5	2
Total issues	3	8	5	8	7	9	8	11	7	6	10	10

Note: † indicates passages shared with main PIRLS test

Figure 7-4 Summary of item functioning and fit per passage

The passage level DRF analysis revealed several passages that functioned differently for Sepedi and Afrikaans focal groups compared to the reference group (English). Based on the unsigned DRF (uDRF) statistics, 11 of the 12 passages showed significant differences ($p < .001$) between the absolute values of expected scores for Sepedi students relative to the English group. Passage A was the only passage where the occurrence of DIF (3 DIF items) did not have a significant overall impact on the way in which the passage functions across the Sepedi and English groups. For Afrikaans students, four passages (B, C, E, and G) displayed significantly different uDRF compared to the English group ($p < .001$). The uDRF statistics represent the average absolute difference in expected scores between groups across the distribution of student ability. Smaller values, i.e. closer to 0, indicate the groups have similar expected performance, while higher uDRF values signify greater dissimilarity in expected performance across groups. Table 7-4 shows the sDRF and uDRF statistics for each passage that functioned significantly differently across reference and focal groups (p < .001) (sDRF statistics for all passages available in Appendix E-1). Also included in the table are the results of the likelihood ratio tests for significance and the confidence intervals for each DRF statistic.

Table 7-4

Passage-level differential response functioning statistics by focal group

Passage	Items	Expected score	Statistic	DRF	5% CI	95% CI	χ^2	df	p-value
Sepedi									
B	15	15	uDRF	.84	.51	1.25	27	2	< .001*
C	13	14	uDRF	.71	.23	1.13	11.71	2	.003*
D	12	13	uDRF	.55	.27	1.04	11.29	2	.004*
E†	15	15	uDRF	.48	.29	.84	14.51	2	< .001*
F	13	14	uDRF	.82	.47	1.36	16.78	2	< .001*
G†	13	14	sDRF	.66	.24	1.01	1.84	1	< .001*
			uDRF	.66	.31	1.02	13.22	2	.001*
H	15	15	sDRF	-1.13	-1.63	-.67	21.63	1	< .001*
			uDRF	1.29	.86	1.79	56.24	2	< .001*
I	12	13	uDRF	.67	.37	1.21	14.17	2	< .001*
J	18	20	sDRF	-.89	-1.35	-.45	15.58	1	< .001*
			uDRF	.9	.56	1.38	42.34	2	< .001*
K†	8	8	sDRF	-.67	-.95	-.37	2.98	1	< .001*
			uDRF	.73	.45	1.03	48.1	2	< .001*
L†	4	4	uDRF	.21	.11	.35	13.21	2	.001*
Afrikaans									
B	15	15	sDRF	.62	.21	.99	9.44	1	.002*
			uDRF	.63	.28	1	11.64	2	.003*
C	13	14	uDRF	.71	.39	1.02	21.35	2	< .001*
E†	15	15	sDRF	.67	.33	1	15.68	1	< .001*
			uDRF	.67	.37	1	19.34	2	< .001*
G†	13	14	sDRF	-.57	-.9	-.22	11.92	1	< .001*
			uDRF	.57	.24	.89	24.68	2	< .001*

*Notes: * indicates χ^2 test was significant at p < .01*
† indicates passages shared with main PIRLS test

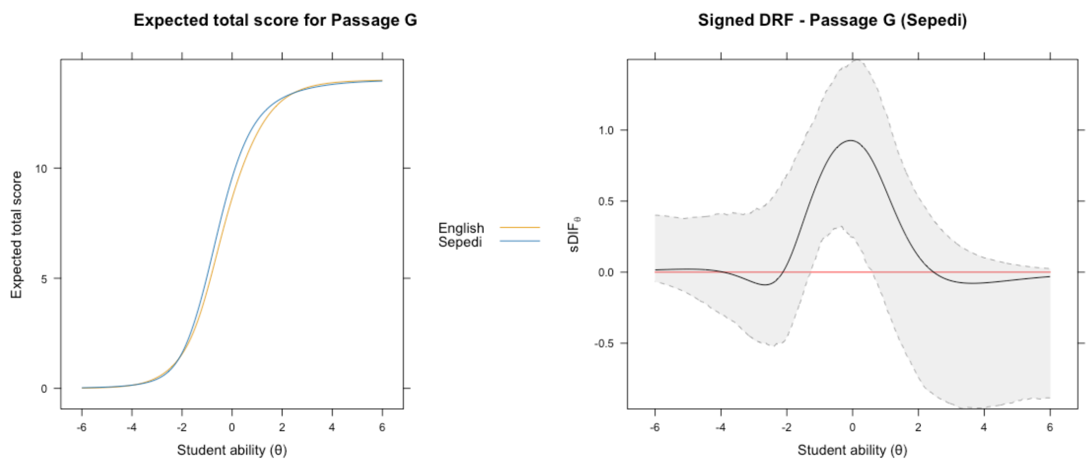
DRF statistics are in the metric of the total expected score, which means seemingly smaller uDRF scores, for instance Passage E (uDRF = .48) and Passage L (uDRF = .21), need to be considered in the context of the maximum scores to understand whether they indicate more modest divergence in score. Passage L, for instance has the lowest uDRF but due to a high number of misfitting items removed during the fit checks (see Figure 7-4) it also has a lower maximum score (max = 4) compared to other passages. So while the uDRF values of .21 to 1.29 are statistically significantly different from 0 based on p-values, their magnitude relative to maximum passage gives an indication of the extent of their impact on scores across groups. To understand if these absolute differences advantage one particular group, signed DRF statistics were considered.

Several passages showed significant sDRF statistics indicating systematic score differences that disadvantage one group relative to students of similar ability in the other group. In other words, students from the one group are likely to achieve a lower expected score for the passage when compared to students with the same ability levels in the other group. Four passages (G, H, J, K) had significant sDRF statistics with Sepedi as a focal group. Passage H again had the largest difference between groups with sDRF = -1.13, 95% CI (-1.63, -.67). The negative values of sDRF show an advantage to the reference group. Three of the four passages with significant sDRF for the Sepedi group were negatively signed, indicating that the direction of the difference in Passages H, J and K benefits English students.

For the Afrikaans focal group, Passages B, E, and G displayed significant differential functioning relative to the English reference group based on sDRF statistics. Passage E had a positive sDRF of .67 (95% CI .33 to 1), which means that this passage was consistently easier for Afrikaans students compared to English students of matched proficiency. By contrast, Passage G favoured English students, with a negative sDRF of -.57, 95% CI (-.9, -.22) indicating that the impact of DIF occurring in this passage (4 DIF items) resulted in the passage presenting a higher level of difficulty for Afrikaans students.

The significant uDRF and sDRF statistics provide evidence that numerous passages functioned differently for Sepedi and Afrikaans focal groups compared to English students of equivalent proficiency. Figure 7-5 displays the expected total score functions (left) and signed DRF plots (right) for the four passages that exhibited significant differential functioning in both absolute and directional terms between the Sepedi and English groups. The expected score plots show the total passage scores expected across the ability distribution for each group. The signed DRF plots illustrate the direction of advantage, with values above the red line indicating regions where Sepedi students are expected to achieve higher scores than English students along the range of student ability, and vice versa.

The sDRF curves for Passages H, J and K show that the differential functioning in these passages advantages different groups across different levels of ability. In all three passages, in the lower ability range $\theta < -1$, Sepedi students are advantaged, while English students are advantaged in the higher ability range $\theta > -1$. This is further evidenced in the expected total score plots on the left which show a crossing point around $\theta = -1$ in each passage where the alternative group is advantaged. In all three cases, the sDRF values were negatively signed indicating that, although the direction of advantage changed at different ability locations, on average across the full range of ability these passages disadvantaged Sepedi students. Passage G, on the other hand, shows an advantage to Sepedi students in the range $-2 < \theta < 2.3$ as well as an average advantage across the full ability range investigated.



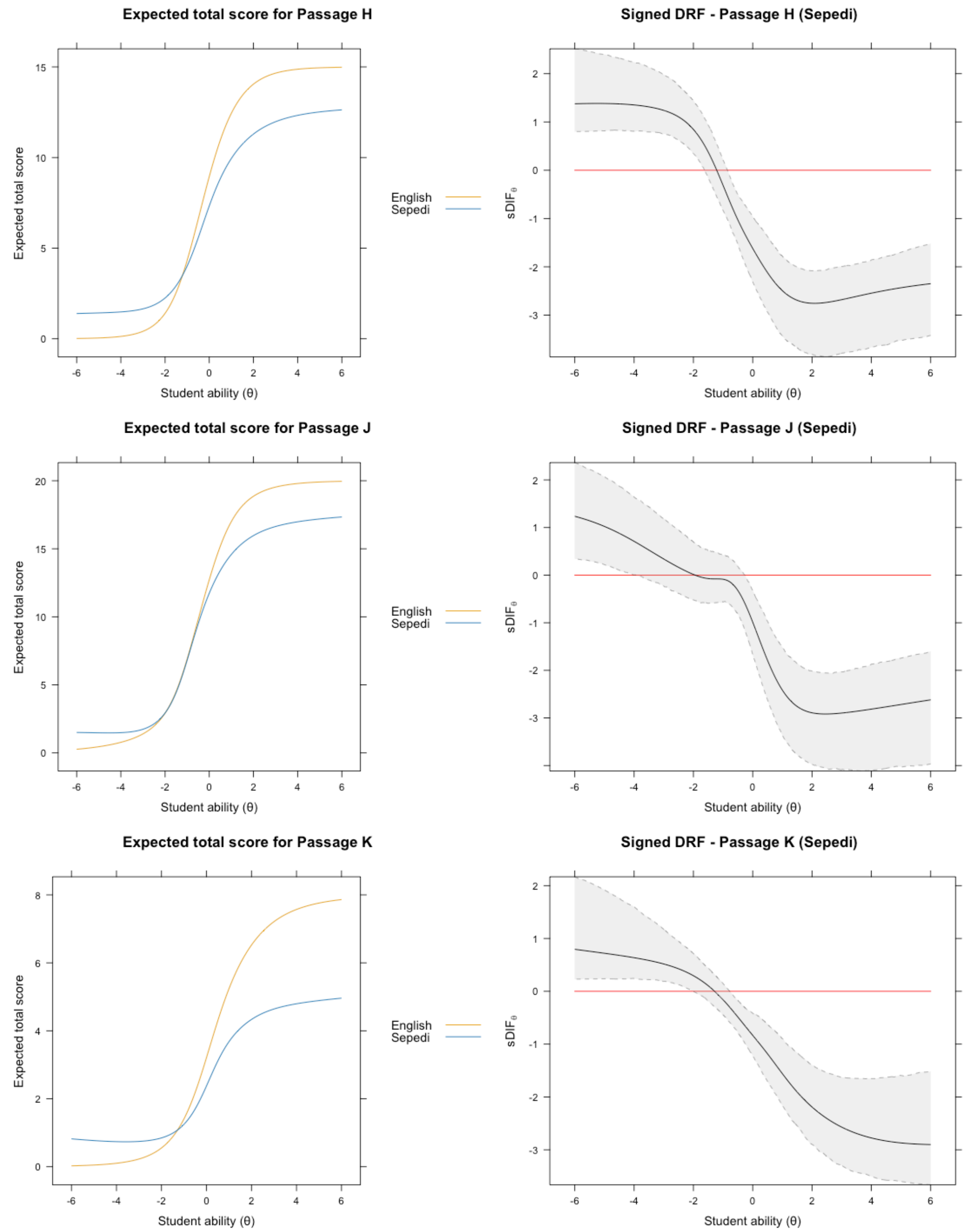


Figure 7-5 Differential response functioning plots for Sepedi focal group – Passages G, H, J and K

Figure 7-6 shows the expected total score functions and signed DRF plots for passages that showed directional differences for the Afrikaans focal group (B, E, and G). In Passage G, the negative sDRF line indicates this passage favoured English students across the ability range examined. Passage B and E, however, showed significant differential functioning favouring Afrikaans students for at least a portion of the ability range. Although Passage E reveals slight advantage English students at lower proficiency, and Passage B shows a crossing pattern indicative of some non-uniform DIF, the broad confidence intervals (indicated in grey) reveal that there is less substantial directional bias for Afrikaans students overall. This contrasts with the Sepedi focal analysis where several passages showed clear changes in advantage at different ability levels.

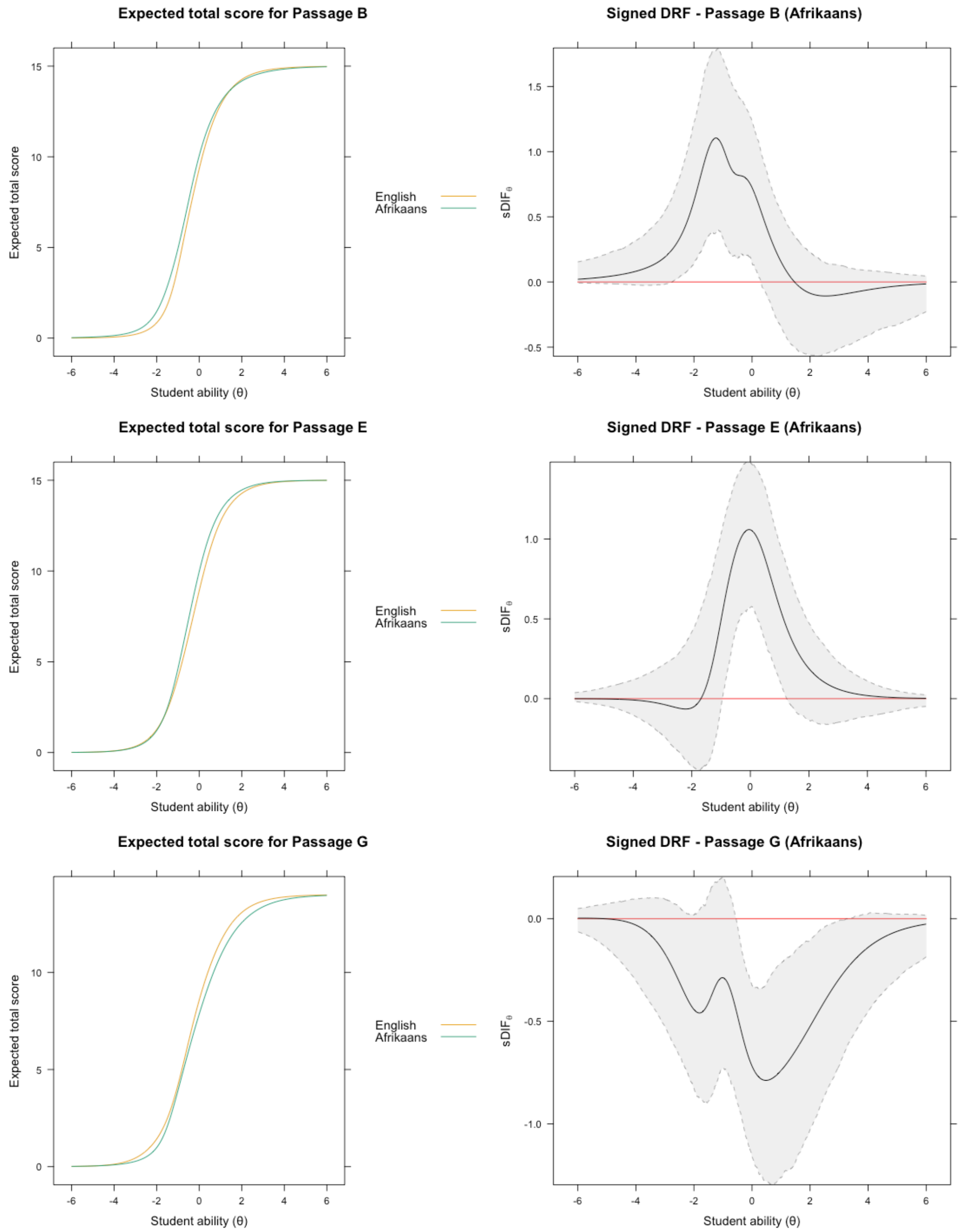


Figure 7-6 Differential response functioning plots for Afrikaans focal group – Passages B, E and G

7.5.4 Evaluating differential item functioning impact at test level

Establishing comparability at an overall test level was undertaken using the same approach as the passage level analysis. However, in the test level analysis, the DIF impact across all items in the test was investigated to determine if the cumulative impact of differential functioning at the item level resulted in significant bias overall. The uDRF and sDRF statistics at a test level for each focal group comparison are shown in Table 7-5.

Table 7-5

Test-level differential response functioning statistics by focal group

Focal group	Statistic	Total items	DRF	5% CI	95% CI	χ^2	df	p-value
Sepedi	sDRF	152	-1.24	-3.7	.81	1.19	1	.275
	uDRF	152	5.29	4.15	7.26	72.97	2	< .001*
Afrikaans	sDRF	152	1.86	-.32	3.89	3.09	1	.079
	uDRF	152	1.93	.54	3.92	5.25	2	.073

*Note: * indicates χ^2 test was significant at $p < .01$*

The results show significant unsigned DRF for the Sepedi focal group compared to the English reference group. Across all 152 items, the Sepedi test had a uDRF of 5.29, 95% CI (4.15, 7.26), indicating substantial absolute divergence in expected total scores between the two language versions. This uDRF effect signifies that on average, Sepedi and English students with equivalent reading proficiency are expected to differ in their total PIRLS Literacy scores by over 5 points. The sDRF of -1.24 (95% CI -3.7, .81) for Sepedi was non-significant, suggesting no consistent directional bias overall favouring either group at the test level despite the passage-level differences. For Afrikaans, neither the uDRF (1.93, 95% CI .54, 3.92) nor sDRF (1.86, 95% CI -.32, 3.89) reached statistical significance at the test level. Figure 7-7 shows the expected test score curves for the Sepedi and Afrikaans groups alongside the sDRF plots[11] for each group at the test level.

[11] Note that the sDRF functions shown in this figure do not represent significant differences between groups

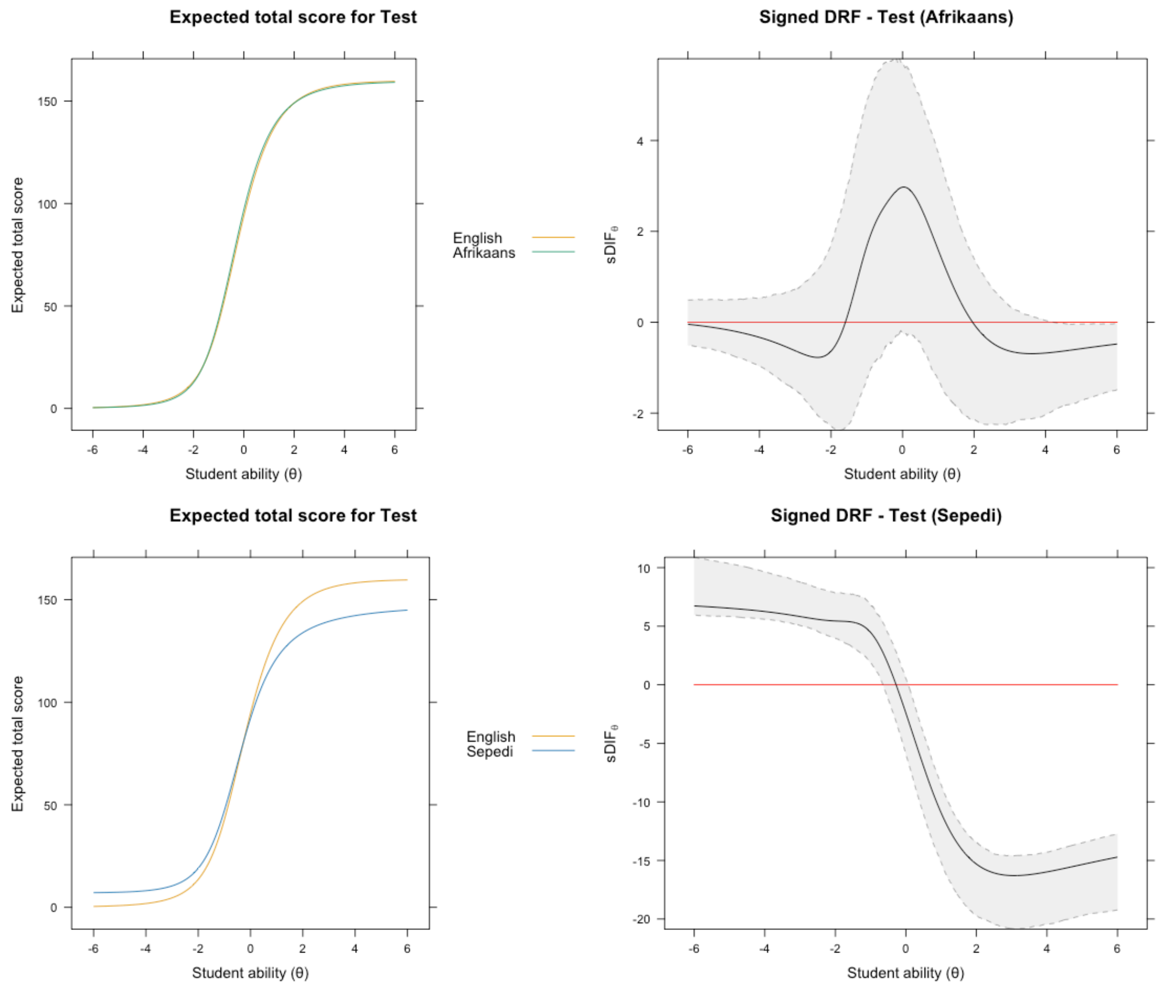


Figure 7-7 Expected test score functions and sDRF plots for both focal groups (n = 152 items)

The non-significant signed DRF (sDRF) alongside the substantial unsigned DRF (uDRF) for the Sepedi test version indicates there may be multidirectional differential functioning that somewhat cancels out in terms of overall directional bias. This pattern is consistent with the presence of non-uniform DIF, where certain items show bias favouring Sepedi students at lower proficiency levels but favouring English students at higher proficiency levels (or vice versa). In such cases, uDRF is a useful consideration as it is able to detect the absolute divergence while sDRF may not reflect consistent directional test-level bias. Therefore, considering both statistics provides vital complementary information regarding comparability when non-uniform differential functioning is present, as was seen in the item-level analysis.

7.6 Discussion

ILSAs like PIRLS are designed to provide valid, reliable, and fair comparisons of student achievement across a wide range of cultural, linguistic, and national contexts. This means that establishing comparability across the various participating groups is an essential step in demonstrating the validity of inferences made from PIRLS results. However, despite rigorous translation and adaptation efforts, lingering issues with comparability across language versions continue to threaten the validity of score interpretations (Ercikan & Por, 2020). This paper evaluated the comparability issues at an item level (RQ1) as well as the impact those issues have at a passage (RQ2) and overall test level (RQ3) of three language versions of PIRLS 2016 in the linguistically diverse context of South Africa. This is a particularly important undertaking due to the low overall performance of students in South Africa, as well as the wide gaps in achievement across language groups. Furthermore, PIRLS results have a high status in South Africa and have been shown to have an influence on policy reform, resource allocation, curriculum reforms, and initiatives to improve reading in the country (DBE, 2023; Mulongo & Amod, 2019; Spaull & Pretorius, 2022). This means that evaluating evidence of systemic comparability issues across language versions is important to ensure that the decisions being made are based on valid information about students reading ability.

Regarding item-level comparability (RQ1), findings revealed that when comparing Sepedi and Afrikaans students to English students with the same overall reading ability, a number of items in PIRLS functioned differently. The differences were more substantial for Sepedi, with over a quarter of items presenting a different level of difficulty than in the English version. The proportions of DIF found across these language versions were similar to those found in previous DIF analyses of PIRLS in South Africa, which looked at passages in the Afrikaans, English, and isiZulu versions (Roux et al., 2022); and English and isiXhosa versions (Mtsatse & van Staden, 2021). This pattern illustrates prevailing evidence of comparability issues found across multiple language versions in South Africa and highlights the importance of investigation into the comparability of language versions that have not yet been investigated.

Another key finding from the item-level analysis relates to the nature of the DIF found. A higher proportion of DIF items displayed non-uniform DIF, which means the direction of advantage found between groups changed across the range of ability. In most instances of non-uniform DIF, students in the Sepedi group were found to have an advantage at lower ability levels but as student ability levels increased, the advantage shifted in favour of the English group. As this pattern mostly occurred in multiple-choice items, the influence of differential guessing behaviour among students with lower overall proficiency across groups is an important consideration.

Differences in guessing behaviour may be related to cultural differences or different levels of engagement in the test leading to students in some groups being more likely to guess the answer than others (DeMars et al., 2013; Svetina Valdivia et al., 2023). In this context, the large gap in overall achievement across groups is an important consideration. The low overall performance of the Sepedi group compared to the English group may have contributed to lower engagement and motivation, potentially leading to more guessing. However, irrespective of the underlying causes, differential guessing across subgroups has been shown to compromise both reliability and validity of test score interpretation, negatively impacting the accuracy of the estimation procedures and undermining fair cross-group comparisons (Deng & Rios, 2022; Guo & Ercikan, 2020).

These observations raise concerns not just about the comparability of item difficulty across groups, but also about the overall effect of poor alignment between student ability and item difficulty for PIRLS results in South Africa. DeMars and Jurich (2015) demonstrated that guessing behaviour can distort DIF findings when there are wide ranges in ability across groups. They underscore the importance of carefully considering the targeting of items to ensure accuracy of the DIF findings. The targeting and potential influence of flooring effects was carefully scrutinised before the DIF analysis was conducted in this paper, and a number of items were removed from consideration in the DIF analysis due to issues with fit and the small proportion of correct responses for some groups. However, the prevalence of non-uniform DIF advantaging low ability students from the lowest performing group suggests that the targeting concerns may be more substantial than initially thought. This has consequences not just for the findings of this study, but also for the potential of other secondary analyses of PIRLS to provide insight into the data more broadly.

The test level comparability (RQ3) analysis across the 152-item set investigated suggests that overall test score comparisons between the English and Sepedi group may be problematic. However, the removal of misfitting items means that these test level comparisons are under representative of the test that students actually took. This is further complicated by the matrix block design of PIRLS where each student only responds to around 30 items. A benefit of this design approach is that it enables broad content coverage in a reasonable amount of time so the students are not heavily burdened (Rutkowski et al., 2013). However, each item that is not appropriately targeted to the students incrementally threatens construct coverage and subsequently may impact the validity of PIRLS for these students.

In light of these constraints, this discussion focuses primarily on the insights that can be derived from passage-level findings rather than the test-level comparisons. Furthermore, the passage-level comparisons also considered the relative proportions of item-level issues occurring in each passage. Comparing the prevalence of these issues across passages can reveal patterns indicative of more systemic measurement challenges for the students being investigated. Added to this, in an instrument like PIRLS, the passages are associated with sets of items and therefore constitute a major part of the assessment design. This means that properly equated passages are an essential component of the overall validity of the test.

The passage-level analysis (RQ2) revealed substantial differential functioning across groups. When compared to English students with the same ability levels, 11 passages functioned differently for Sepedi students, and four functioned differently for Afrikaans students. Directional differences in passage functioning were also considered as these provided evidence not just of comparability issues across passages, but of persistent directional bias across groups. Three passages disadvantaged Sepedi students who achieved expected scores 4-8% lower than students with the same overall ability in the English group; one passage disadvantaged Afrikaans students; and three disadvantaged English students.

This evidence of systematic differential functioning at the passage level has implications for the rotated booklet design of PIRLS, which relies on the premise that passages serve as interchangeable units in order to ensure comparable and comprehensive construct coverage. Although some PIRLS Literacy passages are designed to present a different level of difficulty by design, the relative differences in the number of fit issues found across passages may impact the extent to which the goal of construct coverage is met for all students. The differential impact of DIF found across passages further challenges this design, as passages presented different difficulty levels to students based on the language in which they took the test and not their overall ability level. As students respond to just two passages, such passage-level divergence raises concerns about the validity and the fairness of the rotated booklet design because students' scores become dependent on the particular passages they encountered, as well as the language they took the test in.

Another key finding that emerged from the passage-level investigation relates specifically to the passages that aim to link PIRLS Literacy to the main PIRLS tests. Having an easier version of a large-scale test is one of the methods applied for assessing students with disparate levels of achievement. This approach has been implemented in other ILSAs such as PISA and is recommended by the ITC Guidelines for the assessment of diverse populations (ITC, 2019). Passages designed to link these easier test versions to the main assessments therefore play an integral role in enabling valid score comparisons across countries (Rutkowski et al., 2013). However, traditional equating methods often fail to account for large ability gaps, resulting in errors that undermine score comparisons (Sandefur, 2018).

Two of the passages intended to link the easier PIRLS Literacy test to the main PIRLS assessment displayed the most extensive evidence of issues. In Passage K and L, less than a quarter of items were suitably targeted and comparable across groups. Added to this, Roux et al. (2022) specifically noted that Passage K presented a particularly high level of difficulty for South African students. Crucially, as Michaelides and Haertel (2004) caution, when tests do not adequately overlap in matched items across ability levels, substantial “linking errors” can distort scores between versions. With so few items from the PIRLS linking passages showing comparable and appropriate levels of difficulty, considerable doubts emerge regarding the accuracy of their linkage function. Similar concerns about linkage functioning were raised by Rutkowski and Rutkowski (2021) regarding the items used to link the easier version of PISA (PISA-D) to the main test, which were much too difficult for students across all countries that participated in PISA-D.

The ITC Guidelines clearly state that tests need to consider the range of student abilities appropriately, as well as the variance of the ability distribution between language groups (ITC, 2019, para. 2.7.1). Although PIRLS has provided an easier version of the assessment for countries with low overall performance, the evidence here shows that even the easier version of the assessment does not adequately account for the range of reading abilities across groups in South Africa. This means comparisons between groups are not valid, and fair interpretations on the basis of these scores may not be an accurate reflection of students’ reading achievement.

The impact of the differences in reading ability levels across groups has further implications to the extent of comparability across passages. Although directional bias was seen for some passages, the high prevalence of non-uniform DIF can obscure directional bias that might be occurring due to the potential for cancellation effects (Temel, 2023). This is further compounded by the wide range in achievement across language groups in South Africa which may lead to an increase in guessing behaviour for students from lower-ability groups. To account for this, the absolute differences in expected scores were considered at the passage level. Absolute differences were seen in all but one of the passages when comparing Sepedi and English students. This means that the actual differences between achievement for students from these groups is being obscured by cancellation effects across the ability range. In other words, if there were no DIF items found, differences in achievement seen in the overall reading scores in PIRLS have the potential to be even more substantial.

Typically, the next step in analyses such as this would be to explore why these differences occurred to inform future test improvements. A number of researchers have taken this approach to try to understand the root causes of DIF in ILSAs (e.g., Allalouf et al., 1999; Finch & Hernández Finch, 2016; Huang et al., 2016; Oliveri et al., 2013). However, that was not the concern of this investigation. The results from PIRLS have already been used to inform policy changes and interventions in South Africa, often relying on past cycles (e.g., Gustafsson, 2019b; Gustafsson & Taylor, 2022; Motshekga, 2021; Spaull, 2022; van Staden et al., 2019; van Staden & Howie, 2014). This study, therefore, prioritises properly interrogating the quality of existing assessments that drive consequential decisions to ensure they are based on valid and fair information. Despite the thorough quality assurance standards applied in the development of PIRLS for international cohorts, this analysis has found pervasive evidence that the test does not present a comparable and valid assessment of reading comprehension for students in South Africa. These findings suggest that PIRLS 2016 does not adequately account for the within-country diversity seen in South Africa and potentially useful insights regarding inequalities in reading achievement are obscured rather than revealed.

The cultural, economic and linguistic diversity within South Africa requires nuanced, context-specific assessments that are more tailored to local contexts if the fundamental requirements of validity and comparability are to be met. As a result, using PIRLS results to highlight between-language achievement gaps or guide policy interventions to support reading development in South Africa may be highly problematic.

7.7 Limitations and future research

While the present analysis reveals considerable issues with the comparability and validity of PIRLS for the assessment of reading in South Africa, some limitations should be noted. Foremost is that the concerns about targeting are likely to confound these findings to some degree, although they nevertheless indicate substantial validity concerns.

There are also methodological challenges as several factors can impact DIF detection and interpretation, including the analytical methods used and the ability distributions of the groups being compared. With different methods or more balanced score distributions, additional insights may have emerged. However, the overall very low achievement levels indicate that substantive DIF was unlikely to impact the conclusion that there remain serious targeting concerns across groups.

Another limitation relates to the potential for other contextual influences to affect cross-language comparability. In South Africa, socioeconomic status is strongly intertwined with language backgrounds and as such the examination of potential contextual influences could deepen understanding of what may be causing the comparability issues seen across language groups. It is also notable that the test language was the medium of instruction at school but not necessarily students' home language. Although home language rates were higher among lower achieving groups, differential exposure to the test language at home merits further consideration in future investigations. Additionally, only three language versions underwent investigation here. Expanding the analysis to the remaining eight languages assessed in South Africa could more fully describe the extent of differential functioning for the country's diverse linguistic landscape.

Finally, as emphasising score-level comparability was the primary focus of this research, other comparability concerns such as item features or translation quality were not explored. This is an important area of research, particularly due to the dearth of research regarding reading development in African languages. However, although more granular analyses of translation quality and item features would expand understanding further, this investigation underscores the impact that differential functioning can have on overall comparability across versions. Even with potential confounding factors like low overall performance and limited ability distribution variability, the systematic analysis provides strong evidence of substantial comparability issues that cannot be disregarded.

While the limitations outlined above are important considerations, this psychometric investigation of PIRLS across languages in South Africa offers critical insights that can inform future efforts to enhance fairness and validity in large-scale assessments. Though only three language versions were examined here, this paper highlights the need for similar scrutiny across all languages used in PIRLS administration at both within and across country levels. Moreover, explicitly grappling with the implications of extremely low overall achievement is an essential component of understanding the challenges that are specific to developing country contexts.

CHAPTER 8 DISCUSSION

This research investigated PIRLS Literacy 2016 in South Africa to understand what can be learned from PIRLS data about the reading achievement of South African students across different languages. PIRLS results carry high status in South Africa, informing educational policies, resource allocation, reading initiatives, and progress towards national development goals (DBE, 2023; Spaull & Pretorius, 2022). However, given the linguistic diversity, cultural complexity, and vast inequality characterising South Africa's educational landscape, dedicated analysis of the validity and comparability of PIRLS 2016 is essential before relying on the data for consequential decisions.

Drawing on foundational principles of quality assessment in diverse contexts, this thesis analysed the South African PIRLS 2016 dataset through three interconnected papers. Paper 1 explored relationships between contextual factors, language use, and reading achievement, revealing that the language of the test played a considerable role in students' achievement in South Africa. Paper 2 investigated the extent to which PIRLS is fit for the purpose of assessing reading across groups in South Africa, revealing violations of fundamental measurement assumptions and poor targeting across groups. The analysis in Paper 2 raised substantial validity concerns about PIRLS in South Africa, particularly for the purpose of informing targeted reading interventions. Finally, Paper 3 evaluated the comparability of PIRLS across language versions, uncovering evidence of concerns that impact the validity of score comparisons across groups. Paper 3 also revealed potential issues with the linking design between the main PIRLS assessment and the easier PIRLS Literacy version.

Together, the findings across all three papers present a nuanced picture of PIRLS as a reading assessment for South Africa. The detailed contextual analysis in Paper 1 offers insights for teachers, researchers, and policymakers; underscoring the significant role of language of instruction for reading development in South Africa. Although the concerns uncovered in the subsequent psychometric evaluation in Papers 2 and 3 suggest that PIRLS results should be interpreted with careful consideration in South Africa, the substantial gap in achievement based on test language observed in Paper 1 remains valuable for highlighting the importance of finding strategies to improve the inequality in reading across language groups. The detailed findings pertaining to the analyses in each paper have already been discussed within the papers themselves. This overarching discussion chapter synthesises points of intersection in the findings across the three papers and presents a discussion of broader implications of these findings for assessment in the South African context.

8.1 Missing the mark

Although a number of concerns emerged in the papers, poor targeting stood out as a central issue resonating across all three papers. Specifically, analyses revealed that the overall difficulty level of the test was substantially misaligned with the abilities of the students, despite being designed to be an easier version of the main PIRLS test. Appropriate targeting is essential for quality testing; when a test is too difficult students are unable to demonstrate their abilities on the construct, compromising the meaningful interpretation of scores. The sections that follow discuss the threats that these targeting problems present to the fundamental pillars of validity, reliability, and fairness for PIRLS in South Africa.

8.1.1 Establishing validity

As described in Section 3.2, a core concern of validity is how well a test assesses the construct it aims to measure. This can be difficult for multifaceted constructs like reading ability, which encompass a number of different skills and processes. PIRLS recognises this complexity in its reading framework and includes multiple items and passages to capture the different reading purposes (literary and informational) and comprehension processes (e.g. retrieving directly stated details, interpreting and integrating ideas). To ensure comprehensive construct representation, PIRLS also employs a matrix design technique so connections can be made across student performance on different item sets, even if individual students do not respond to all items. However, the severe targeting issues seen in Paper 2 and Paper 3, raise concern about the validity of the construct coverage for South African students. If all students are not able to comprehensively demonstrate their ability on the different components of the construct, then the fundamental expectation that the assessment accurately reflects students' reading ability is compromised.

Validity encompasses more than just construct coverage, ensuring a valid test also involves considering the social consequences of score interpretations as well as the extent to which the test appropriately serves the intended purposes within its applied contexts (Messick, 1998). When evaluating validity, therefore, it is necessary to consider the social consequences connected to the meaningful interpretation and use of the test scores. PIRLS is used for two primary purposes in South Africa: monitoring progress towards national reading development goals; and informing education policy and targeted resource allocation (DBE, 2023; SAHRC, 2021).

To enhance the quality of information provided and make inferences drawn from the results more meaningful, PIRLS provides a set of four international benchmarks. These benchmarks provide criteria that describe overall achievement in line with the different purposes and processes that form part of the construct of reading. The new national reading assessment framework in South Africa directly links PIRLS benchmark criteria to the national development goal – ensuring all 10-year-olds can read for meaning by 2030 (see Section 2.4.2). The consequences of poor targeting therefore, mean that extent to which PIRLS data is able to provide valid metrics to track the progress towards that goal is compromised. Added to this, flooring effects resulting from the misalignment between test difficulty and student ability means that the discriminative value of the test is reduced, limiting the capacity of the assessment to provide detail on the variation in reading abilities.

While acknowledging the challenges with the overall low performance, South African policymakers highlight the potential utility of PIRLS "microdata" to provide nuanced insights into reading comprehension across subgroups for informing targeted interventions and resource allocation to most needed areas (DBE, 2023). Paper 1 demonstrates the potential utility of this data by revealing that test language strongly influenced student achievement even when accounting for other potentially influential factors, particularly for students taking the test in one of the nine

African language options. This is an important finding that highlights some of the extreme inequality still facing different groups within South Africa that has led researchers and activists to call for increased attention to the development of resources focused on teaching African languages in South Africa (Makalela, 2023; Mohohlwane, 2019; Motala & Carel, 2019). Moreover, the strength of the relationship between reading achievement in South Africa and contextual and language variables demonstrated in the regression analysis emphasises the importance of considering the influence of contextual and language factors in reading achievement.

Considering the role of contextual factors is a particularly important undertaking in the South African context where historical factors continue to disadvantage some groups more than others (see Section 2.4.3). However, the targeting issues as well as concerns about contextual relevance revealed in this thesis can lead to problems when using PIRLS data to inform interventions or examine achievement gaps. This means that, while PIRLS is able to offer important insights into reading achievement and relevant contextual factors in South Africa, its application should be considered with an awareness of its limitations.

The substantial differences in PIRLS performance based on test language seen in Paper 1 also raised questions about the cross-language comparability of the PIRLS instrument, which motivated the psychometric evaluations undertaken in Papers 2 and 3 revealing that the targeting problems were more severe for Sepedi students. As a result of the disparities in difficulty faced by students from different groups, using PIRLS results to evaluate gaps across language groups may risk misrepresenting their relative achievement to some extent. This is a particular concern as gaps across language groups have become one of the primary focus areas for reading interventions in South Africa, making it crucial to ensure that between-group comparisons are valid and meaningful.

The challenge of interpreting achievement gaps in South Africa is further complicated as it requires grappling with the intricate interplay between language, race, and ongoing socioeconomic inequality (see Section 2.2). Although there are comprehensive background questionnaire variables in PIRLS, the items included in the questionnaires do not adequately capture the unique complexity of the situation in South Africa particularly in terms of race and schooling. While the findings from Paper 1 show the impact of language of instruction on PIRLS results using PIRLS background variables to control for school, home, and student factors; these findings contrast with evidence from robust local studies using more contextually nuanced data. For instance, Taylor and von Fintel (2016) show that when context-specific data that controls for schooling quality constraints, race, and disadvantaged background is used, students who receive instruction in an African language for the first three years perform better later on than those instructed in English first. These contrasting findings show the constraints of using data from ILSAs that are unable to fully account for nuances of certain contexts.

PIRLS tries to account for contextual differences by enabling NRCs to make adaptations to the background questionnaires to better suit the individual needs of their context (Foy, 2018b). For example, to better address the difference types of schools in South Africa, the national questionnaire had an additional response option "township near urban areas" for school location. Townships are unique features of South African geography that emerged under the segregatory apartheid policies, which designated living areas based on race. Specifically, townships were zones designated for black, coloured or Indian South Africans, usually located outside major cities and characterised by poor infrastructure, and inferior living conditions. In contrast, suburban areas with better amenities and larger houses were restricted to the white population.

Although the PIRLS NRC in South Africa expanded the school location options in their national documentation to include township schools, this category was subsequently collapsed with ‘suburban schools’ for the purposes of international comparison (Foy, 2018b). This means that the data available through the public international database does not account for the distinct effects of township and suburban schools. The historical segregation and inequality of apartheid continue to manifest in extreme differences between township and suburban schools today. Consequently, treating these categorically distinct schools as a single entity obscures meaningful distinctions between them.

An additional challenge for secondary researchers is that public ILSA datasets rarely include race or ethnicity data. In the absence of race-specific data, therefore, secondary researchers typically leverage the overlap between race and language in South African schools and use data about students' medium of instruction as a proxy for race (e.g., Taylor & Yu, 2009). This approach risks misclassifying students at disadvantaged schools that recently changed their language of instruction and misrepresenting the increasing within-school diversity in historically white schools (Taylor & von Fintel, 2016). Moreover, using the test language as proxy also risks conflating the enduring effects of historically discriminatory policies with current language-in-education policies, which aim to address these historical inequities. The complexity of the contextual inequalities means that the benefits of home language instruction should not be dismissed without adequate and contextually valid controls for school quality, socioeconomic status, and race (Mohohlwane, 2019).

Acknowledging these complexities, it is important to emphasise that persistently low achievement in African language groups remains a substantial challenge in South Africa, one that has been consistently identified in numerous local, national, and international studies (see Section 2.3). Improving resources for reading instruction in African languages, therefore, continues to be an undisputed priority. By testing in all 11 languages, policymakers in South Africa note that PIRLS has the potential to provide crucial evidence to help inform initiatives aimed at addressing language development needs (DBE, 2023). Indeed, the findings from Paper 1 highlight just how substantial the differences in performance across language groups may be. However, fully understanding the situation requires effectively capturing complex local dynamics so that detailed and meaningful insights into the achievement comparisons across groups can be provided.

While it is understandable for an international study like PIRLS to gather information applicable to the majority of participating countries, tension arises between attaining cross-country comparability and maintaining validity across contexts. This tension increases as the range of participating countries expands, particularly as the widening range of ability has implications for the reliability of the information for countries performing at the extremes.

8.1.2 Evaluating reliability

Reliability refers to the consistency and accuracy of measurement procedures and resulting achievement estimates. If assessment items are too difficult for the students tested, it can introduce floor effects, eroding the quality of measurement and compromising reliability (Foy et al., 2010; Gustafsson & Barakat, 2023; Rutkowski et al., 2019). The consequences of flooring effects are a particular threat for South Africa where the overall achievement in PIRLS has been the lowest of all participating countries in every cycle. As described in Section 2.4, targeting PIRLS items at an appropriate level for the low performing South African sample has proved an ongoing challenge across all previous cycles of participation. The findings from all three papers in this study reveal evidence of similar reliability concerns for the 2016 cycle of participation.

A primary reason South Africa participates in PIRLS is to use achievement data as formal indicators to track progress towards national development goals (SAHRC, 2021). Persistent reliability concerns related to targeting issues across consecutive PIRLS cycles threatens both the accuracy and consistency of the data for monitoring progress over time. Notably, issues with the scaling of the 2011 prePIRLS results to the main PIRLS test have already led to the classification of the 2011-2016 trend for South Africa as unreliable by the IEA (see Section 2.4). Echoing the challenges faced with scaling in 2011, Paper 3 exposed concerns about the passages used to link the easier version of PIRLS 2016 to the main test. This provides further evidence of the longstanding doubts about the reliability of PIRLS data for tracking reading progress in South Africa.

In an effort to enhance the accuracy of measuring reading across wide proficiency ranges, the easier version of PIRLS was discontinued in 2021 and replaced by a multiple group adaptive design, where countries could choose ratios of easier and difficult passages based on past performance. However, major reliability issues persisted for South Africa in PIRLS 2021 (Mullis et al., 2023). Despite administering the largest ratio of easy items permitted, South Africa had less than 30% correct on over half the items in 2021 (130 out of 294 total items), and some items were answered correctly by no students at all (Bristol et al., 2023).

The imprecision that results from floor effects carries substantial implications in the South African context as another core goal of participating in PIRLS is to provide metrics to identify highest need areas so as to target resources towards addressing inequality. However, in the presence of floor effects, achievement estimates lack accuracy, particularly for very low performing groups most in need of support. For example, findings from Paper 2 identified items where guessing was a better predictor of success than overall reading ability level. Pervasive guessing behaviour such as this, even in the absence of floor effects, would be a major threat to the reliability of the assessment instrument (Burton, 2001; DeMars & Jurich, 2015). If students are responding randomly, or using strategies unrelated to their reading ability, their scores will be inconsistent and would not reflect a reliable measure of their reading skills. These impacts were more apparent in the Sepedi group, suggesting that there are items that are not providing reliable information about reading ability for students from some groups. This calls into question the utility of PIRLS data for reliably distinguishing groups most in need of support. Allocating support based on unreliable estimates may overlook critically underperforming areas or disadvantage already vulnerable populations.

Although these reliability concerns raise questions about the precision of PIRLS data for identifying and understanding detailed information about the lowest-performing groups, it is important to note that similar patterns of achievement are regularly observed across language groups in South Africa. The findings from Paper 1, along with other research into student outcomes in South Africa (e.g. Spaull & Jansen, 2019; Taylor & von Fintel, 2016; van Staden & Combrinck, 2023), have repeatedly demonstrated a substantial achievement gap between students in English and Afrikaans groups compared to African language groups. This persistent trend, despite the measurement challenges, underscores a clear need to understand and address the factors that affect the reading performance of students learning in African languages, while also raising pivotal questions about the fairness of the assessment for all participating students.

8.1.3 Considering fairness

Fairness is a central component of quality in ILSAs, not only in terms of the test administration procedures but also considering whether the social consequences of the test scores are equitable (Camilli, 2013; Nisbet, 2019; Ortiz et al., 2018). Given South Africa's overall low performance, a high proportion of difficult test items is expected, but having too many inaccessible items can have adverse consequences for the students taking the test. When students consistently face overly challenging test content, it can cultivate feelings of frustration, anxiety, and low academic self-concept (Bieleke et al., 2023; Fulmer & Tulis, 2013; Wolf et al., 1995). These experiences can also undermine motivation and engagement, and students who struggle on assessments may start to withdraw effort, believing achievement lies out of reach (Eklöf, 2010; Wise & DeMars, 2010). Over time, low levels of motivation and self-competency can curtail learning for students, further widening gaps across high and low performing groups (Chapman et al., 2000; Pekrun, 2006). Moreover, early experiences of frustration in reading can inhibit confidence in reading, which in turn affects later reading proficiency and subsequent academic achievement across all subjects (Chapman et al., 2000; Fulmer & Tulis, 2013).

Findings from Paper 1 underscore this relationship, revealing that confidence in reading was strongly associated with higher achievement in PIRLS for students in South Africa. The impact of overly difficult test content on students, therefore, raises critical questions around the fairness of the assessment, as participating in PIRLS may be placing unfair strain on students. This is particularly problematic because findings in Paper 3 revealed that some students found the test more challenging based on characteristics of group membership and not their reading ability. Added to this, because negative testing experiences can have adverse long-term impacts on academic performance, the imbalance in difficulty for students across groups may be further perpetuating inequality gaps. This is a concerning consequence for South Africa where PIRLS participation is, in part, motivated by the desire to improve inequality that persists in the educational system (see DBE, 2023).

Importantly, as Nisbet and Shaw (2019) highlight, the consequences of an unfair test extend beyond just the impact for students taking the test. Because important policy decisions aimed at improving educational outcomes are made on the basis of PIRLS results in South Africa, an unfair test could lead to misguided strategies that negatively impact not only future students, but also teachers, policymakers, and government initiatives working to address inequality in the education system. For instance, while the prioritisation of reading development is undoubtedly crucial for educational improvement, the misalignment between the difficulty of PIRLS and the ability of South African students raises critical questions about the appropriateness of using PIRLS benchmarks to guide local reading programmes.

The South African national Curriculum and Assessment Policy Statement (CAPS) states that Grade 4 reading comprehension activities should utilise texts between 150-200 words (DBE, 2011b). By comparison, the passages in PIRLS Literacy 2016 were between 400 and 800 words. According to CAPS, students are not expected to engage with texts of 400-500 words until Grade 7 (12-13 years old) (DBE, 2011c). This misalignment between the expectations of PIRLS and the expectations of the national curriculum could also have unintended negative consequences for teachers. The low performance of students in PIRLS may be misattributed to problems with teachers' classroom practices rather than fundamental differences in assessment-curriculum alignment. The use of the PIRLS framework, therefore, risks adding pressure to increase expectations in the South African curriculum rather than scaffolding developmentally appropriate steps that align with students' current skills and abilities.

Furthermore, if pressure to meet unrealistic PIRLS expectations becomes an educational priority, this could undermine efforts to establish essential reading foundations first. For instance, underscoring "reading for meaning" as a primary instructional focus and national development goal – while certainly an important long-term objective – risks minimising instructional emphasis on foundational decoding and language skills necessary for reading development. Although some reading initiatives in South Africa do address these foundational components (e.g., EGRS), the prominent role of PIRLS could inadvertently lead teachers to prioritise higher-order skills before core foundational skills like phonemic awareness, decoding, and vocabulary are firmly established. Moving forward, incorporating realistic and progressive benchmarking aligned to students' developmental readiness across the early grades may better support equitable reading interventions in the South African context, while also ensuring assessments do not subject students or teachers to unreasonable expectations or unfair testing experiences.

8.2 Mapping the future

The concerns about the validity, reliability, and fairness of PIRLS in South Africa raise critical questions about the value and appropriateness of continued participation. Despite these challenges, reflecting on South Africa's past engagement in PIRLS (as well as other large-scale assessments), can help to reveal valuable insights and shed light on important considerations for the future of large-scale assessments in the country.

8.2.1 Lessons learned

As discussed throughout this chapter, the main problem arising is that PIRLS has been, and remains, too difficult for South African students. Although this means that the quality of detailed insights and comparisons from the data is compromised, it does give some indication that the reading levels of South African students are concerning relative to international comparators. Added to this, because the overall ranking of South Africa has remained very low for every cycle, there is reasonable (although not robust) evidence to suggest that the concerning reading levels have not progressed substantially over the past few cycles. It is important to note that this does not mean there has been no progress in South African reading, the lack of precise and reliable data about performance over time prevents insights from being drawn about potential progress markers. But precisely because the results have remained at unreliable thresholds from cycle to cycle, indicates that gains have not been substantial enough to meet the expectations of PIRLS difficulty levels.

Although extracting crude cross-country rankings from the PIRLS data can be problematic, these high-level comparisons often steer prioritisation and strategic decisions to guide reform (Lin et al., 2014). As Sandefur (2018) explains, global organisations frequently use surface-level rankings from ILSAs to determine investment priorities and track progress towards international development goals. Therefore, despite the concerns, South Africa's persistently low performance provides a rough metric to justify that reading development requires ongoing resource prioritisation and improvement in the country. Moreover, although the PIRLS data in South Africa may lack granular detail about the lower-performing African language groups, the overall pattern revealing that students taking the test in African languages have lower achievement in reading remains consistent across multiple cycles and studies. This consistency suggests that the broader insights that can be gained about systemic inequalities, as seen in Paper 1, remains valid and actionable. Therefore, while acknowledging the limitations in the precision of the data found in Papers 2 and 3, policymakers and educators can still use this information to guide intervention strategies at a broader level.

However, after multiple cycles have already demonstrated these challenges, the value of ongoing participation solely for surface-level signalling may risk adverse consequences. First, subjecting students to difficult tests with unrealistic expectations undermines fairness and has potentially demoralising consequences without constructively informing change. Second, since PIRLS has not provided reliable evidence across cycles, the data cannot track growth over time. Finally, and crucially, PIRLS offers little specific diagnostic value towards addressing South Africa's underlying gaps but has the potential to misguide resources and mislead policy reforms. In other words, when considering future participation options, it is not just about what continued participation can offer (other than restating already known reading problems), it is also about considering the potential for detrimental consequences based on participation.

South African policymakers recognise the challenges with PIRLS and have recommended further explorations to illuminate details about potential small-scale reading improvements (DBE, 2023; Gustafsson & Taylor, 2023). Although the findings from papers in this thesis have shown that even smaller-scale analyses may have inherent validity challenges, it is still important to recognise that the rigorous design and framework of PIRLS can be valuable for informing future assessment development in several ways.

Notably, the extensive contextual data collected alongside PIRLS enables in-depth analyses like those undertaken in Paper 1. Such detailed contextual information, when combined with careful interpretation of achievement data, can guide targeted interventions and policy decisions. Although there are often concerns that large-scale studies may not fully capture the unique aspects of local contexts, the methodological approach used in Paper 1 demonstrates how contextual data collected alongside achievement data can be leveraged to gain insight into the situation and to flag important areas of concern about contextual influences on overall performance. Additionally, undertaking contextual analyses of large-scale data helps to develop insights into areas where the questionnaire items can be further adapted or enhanced. This process is valuable as it not only provides useful findings from the available data but also highlights gaps in the current instruments. Identifying these gaps ultimately informs improvements to future questionnaires, allowing for more nuanced and context-specific data collection.

Another area that could be leveraged is examining passages which did function well for South African students, particularly those items with sufficient proportions of correct responses and minimal evidence of guessing. Typically, these were the easier items in the PIRLS Literacy passages that were not shared with the main PIRLS test. A caution here is that PIRLS' matrix-based design means only subsets of students respond to each passage so examining only well-functioning items comes at the costs of generalisability and satisfactory construct coverage. Nevertheless, these items have value for informing future test design by providing insights on accessible content that matches student ability.

Added to this, capacity training for item design and psychometric analysis of data is a key benefit of participation in ILSAs, particularly for developing countries. As Lockheed & Wagemaker

(2013) note, involvement in ILSAs exposes participants to best practices and hands-on learning in areas such as test development, analysis, and quality control. Through working with international experts, education officials from developing countries can enhance their technical knowledge and skills to carry out high-quality, policy-relevant assessments domestically. This capacity, in turn, enables the use of data to inform policies aimed at improving student learning outcomes over the long term. The feasibility of gaining capacity for assessment analysis and development from ILSA participation has been demonstrated through the integration of items and procedures from ILSAs into regional assessments. SACMEQ for example, has used items from TIMSS and PIRLS in previous cycles to enable wider international comparisons, as well as to help guide quality assessment development (Cresswell et al., 2015).

Additionally, although the instruments themselves prove inaccessible for most South African students currently, inspecting the theoretical framework and approaches for understanding reading development can provide opportunities to improve approaches to teaching reading in South Africa. For instance, Taylor (2023) unpacks the PIRLS framework to help inform reading pedagogy for teacher education programmes in South Africa. Several researchers have undertaken similar studies that explore and use the constructs, frameworks, and released passages from previous cycles of PIRLS to better understand what is happening in classrooms in South Africa (e.g., Pretorius & Roux, 2023; van Staden & Howie, 2014; Zimmerman & Smit, 2014, 2016).

Approaches to benchmarking reading achievement using criteria from PIRLS benchmarks have also been used to inform reading development practices in South Africa (Pretorius & Roux, 2023; SAHRC, 2021). Although care needs to be taken to ensure alignment with curriculum expectations, detailed benchmarking criteria can provide a useful metric for scaffolding, guiding, and tracking reading development over time and cohorts. For example, seeking to provide a shared understanding of what reading success looks like in different languages, Ardington et al. (2021) used data from EGRA across three African languages to identify language-specific fluency thresholds necessary for benchmarking early grade reading. These language- and context-specific benchmarks help contribute to guiding curriculum reform, improving literacy progress tracking, and addressing gaps for reading in African languages.

Qualitative insights derived from participation in ILSAs can also enhance and develop an understanding of reading assessment and reading instruction. Teacher feedback on exam content accessibility, student reflections on engagement factors, and expert analysis around comparability limitations all provide crucial contextual perspectives. For instance, Zimmerman and Smit (2016) conducted a comprehensive case study on teachers' reading practices using a mixed-method approach, including investigation of PIRLS 2006 questionnaire data, teacher reviews of PIRLS released passages, as well as semi-structured interviews and classroom observations. By understanding stakeholder experiences, critical diagnostic input can be gained for strengthening the alignment and equity of assessments for informing reading practice in classrooms.

Ultimately, while large-scale assessments intend to inform and improve educational outcomes, South Africa’s experience in PIRLS underscores how improperly contextualised instruments can pose risks to participants without adequately informing change. Moving forward, it will be valuable to carefully consider both the value and cost of participation in ILSAs that are not aligned to South African students, and investigating options for more constructive and better targeted approaches that can contribute more fully towards understanding and addressing gaps in reading proficiency.

8.2.2 Future prospects

As South Africa considers paths forward to address reading it is necessary to evaluate the trade-offs around participation in international, regional, and national large-scale assessments. While withdrawing from PIRLS may hold consequences related to international signalling and funding, complementing it with enhanced national and regional assessments could provide a pragmatic option. This section explores possibilities for better targeted approaches that more reliably measure the complex components of reading ability in South Africa. The complexities involved make this an immense undertaking that requires balancing comparability, validity, and utility with the unique and diverse needs of the multilingual context of South Africa.

Instead of opting out of PIRLS entirely, South Africa could explore working with the IEA to improve alignment of PIRLS for the local context. This would need to include ensuring adequate representation of contextual inequalities in the schooling system, incorporating easier items to connect to international data, and leveraging adaptive testing to better target low ability ranges. Importantly, given the influence of the PIRLS contextual factors on reading achievement seen in the contextual analysis in Paper 1, having more localised approach to contextual data collection could enhance the relevance and applicability of findings for the South African education system. In light of this, future educational policy and practice in South Africa would benefit from leveraging the contextual insights provided by PIRLS while also addressing the targeting concerns identified, ultimately working towards more robust and locally tailored assessment approaches. Moreover, the findings stress the need for future research to help to facilitate a more nuanced understanding of the effects of contextual factors and language factors on reading achievement to better inform related policies and interventions.

However, the feasibility of successfully balancing South Africa's diverse within-country needs with IEA's core focus on international comparability is uncertain. The recent 2021 cycle introduced group adaptive models in an attempt to better target the full range of participating countries. However, reservations about reliability for both South Africa and Egypt were raised by the IEA in 2021. This means that questions persist around whether PIRLS, even in an adapted form, can provide valid insights into the reading proficiency of lower performing countries.

Added to this, although overall participation is growing year on year, the majority of IEA participants remain high-performing, well-resourced countries while the number of African countries is low and declining. For instance, in PIRLS only Egypt, Morocco, and South Africa have participated consistently; Botswana joined PIRLS 2011 but did not return in later cycles. Similarly in TIMSS, six

African countries participated in 2003 (out of 51 total) and, although the overall number of participating countries has grown to 72 in 2023, only three African countries remain.

Digitised or computer adaptive testing is recommended as an important future direction for ILSA as it can allow for more flexible, adaptable, and precise measurement that will enable better accuracy for estimation of performance at the top and the bottom of the achievement scale (Rutkowski & Rutkowski, 2022). However, very high-levels of crime, an ever-worsening electricity crisis, extreme connectivity challenges in some areas, and a widespread lack of technological (and even basic) infrastructure in schools makes this a very unrealistic option for South Africa at present (Gustafsson, 2019b). This was evidenced in PIRLS 2016 when, despite considerable effort from schools and administrators, the lack of technical infrastructure in schools meant that an attempt to administer ePIRLS in just a small sample of schools was unsuccessful (Howie et al., 2017). Notably, this is also a concern for the practicality of South Africa's continued participation in PIRLS which from 2026 will only have a digital option (IEA, 2023).

The IEA recognises the challenges with PIRLS for South African students and has recommended two alternatives: testing older students (e.g., Grade 6) in PIRLS, or participating in the IEA's Literacy and Numeracy Assessment (LaNA) (DBE, 2023). Although LaNA is targeted to students in low- and middle-income countries, switching to LaNA instead of PIRLS comes with challenges. As described throughout this thesis, PIRLS has become strongly tied to national development goals and withdrawal after continued cycles of poor results may signal a lack of government commitment to the reading crisis. This concern is intensified by the widespread and high-profile criticism across South African media and public discourse in response to the poor PIRLS results (e.g., Joint Education Trust, 2023; Maimane, 2023; Maynier, 2023).

Another challenge with both of these recommendations is that testing older students using either PIRLS or LaNA, which tests students at the end of primary school (age 12 in South Africa), comes with additional drawbacks. Firstly, the national reading goal has been adapted to track reading progress at age 10, which would mean participation could no longer be used as a means of monitoring progress toward that goal. Secondly, later assessment periods would have additional challenges for informing language in education policy, which is one of the targets for South Africa's participation in PIRLS. An additional concern with assessing older students is that there have been marginal differences between Grade 4, 5 and 6 cohorts in PIRLS results for South African students over the years, suggesting that the problems with targeting may not be addressed by testing older students. Finally, considering the foundational role that reading plays in educational outcomes, later assessment periods might be too late for adequately informing options for improving reading development. The foundational components of reading development are essential for success, and understanding the progress of students in components such as vocabulary development and decoding skills should therefore be prioritised.

Recognising the pivotal nature of early reading competency, some efforts are underway to measure and address these foundational reading skills in South Africa. Smaller-scale provincial studies like the EGRS and localised projects like Funda Wande have begun quantifying and developing benchmarks for measuring vocabulary, phonics, and decoding ability in some of the official South African languages (DBE, 2022; Makaluza, 2020). Both EGRS and Funda Wande also include intervention programmes and teacher training guides to help understand and improve the teaching of reading in African languages (Ardington & Henry, 2022). Robust data on initial skills required for reading could expose gaps needing urgent intervention before the compounding inequality trajectories seen from late primary assessments. However, although these provincial programmes and initiatives are expanding, there remains a deficit in nationally representative data in all 11 languages about these foundational skills.

One pragmatic path forward could involve linking international assessments with enhanced national and regional assessment programmes. This could help to retain international comparability and leading design principles, while enabling adaptation of tests better suited to contextual needs. Sandefur (2018) recommends this as a particularly useful option for African countries as it allows for continued tracking for international policy and goals while also supporting more localised needs.

These approaches have already been applied in SACMEQ, albeit with disputed levels of success (see Howie, 2022). Additionally, provincial testing programmes like Systemic Testing have integrated items and design principles from TIMSS into provincial mathematics testing programmes (DBE, 2017b). Linking national and regional programmes with international assessments can expand the utility of data for equitable policy improvements in South Africa. However, if this is considered at a national level in South Africa, it is essential that linking is considered carefully during the design stages as a high number of well-targeted shared items are necessary for effective and reliable comparisons with international tests (Sandefur, 2018). Moreover, given resource challenges in South Africa, it is important to decide what needs to be prioritised to address educational inequalities. While ILSAs can provide globally comparative measures, is it worth dedicating resources to link national and international studies if that comes at the expense of better targeted tests that address more foundational reading components?

Concerns about the suitability of assessments developed in high income, higher-performing contexts for use in developing countries have persisted for decades (Ahmed et al., 2022; Heyneman & Loxley, 1982; Kamens & Benavot, 2011; Wagner, 2010). Lin et al. (2014) highlights that countries should carefully evaluate the rationale for participation in light of national educational values and goals, assessing whether the information provided by the assessments is sensible, meaningful and applicable to support and enhance the core local priorities.

South Africa has long recognised there is a systemic crisis in education, in which reading stands out as one of the clearest indicators of the extent of the inequality in the system (Spaull & Pretorius, 2019). Without quality education, the enormous wealth and opportunity disparities in South Africa will not change. Given that reading serves as a foundation for future learning, addressing inequality requires a focus on ensuring all students learn to read proficiently. To this end, it is imperative that we take crucial, well-informed, targeted steps to make sure that reading development is effectively supported, particularly for those students currently facing substantial limitations in accessing quality educational opportunities.

Helpfully, there is generally consensus from government, NGOs, policymakers, teachers, and researchers about what needs to be prioritised to improve reading in South Africa (DBE, 2015, 2023; Gustafsson & Taylor, 2022; SAHRC, 2023; Spaull, 2022; van Staden & Combrinck, 2023). Three main priorities are clearly identified:

- improve African language resources for reading;
- improve teacher education to facilitate better pedagogy for teaching reading; and
- reduce inequalities in school resources and infrastructure.

I propose no changes to these priority areas. I simply recommend ensuring that the information we rely on to inform the steps that need to be taken to achieve them is valid, reliable, fair, and takes account of the unique South African context.

CHAPTER 9 REFLECTIONS

This is thesis has evaluated the extent to which PIRLS Literacy 2016 is appropriate for use in the South African context. By critically examining the suitability of PIRLS within the unique South African educational landscape, this thesis not only sheds light on the local context but also offers insights that resonate with the wider discourse on the use of ILSAs in other linguistically diverse, developing nations. The previous chapter provided a discussion of the findings across all three papers in the thesis as they relate to the context of South Africa; this final chapter synthesises the broader contributions across all three papers, reflects on the overarching limitations inherent in such an undertaking, and discusses the implications of the findings from this thesis for future directions of both research and practice more broadly.

9.1 Limitations

The limitations specific to each paper have already been discussed in the relevant chapters of this thesis, however, the extent of the measurement challenges with PIRLS in South Africa that emerged throughout this investigation necessitates a critical reflection on the limitations of the analyses across all three papers. This is particularly true because the approaches used depend on underlying assumptions about measurement quality, acceptable targeting, as well as model and data fit – all of which were found to be lacking.

A central limitation that permeates the analyses in this thesis is that the targeting issues identified in Paper 2 cast doubt on the validity and reliability of inferences drawn from the PIRLS data. While these measurement challenges introduce uncertainty in the achievement scores used in Paper 1, the findings of wide performance gaps across language groups, particularly the disadvantage for African language groups, remain a critical concern. The magnitude of these gaps is so substantial that, despite reliability issues, they continue to highlight significant achievement disparities in the South African education system. This conclusion is further reinforced by the consistency of these findings across multiple PIRLS cycles, as well as similar observations in the results of other achievement tests in South Africa. Thus, while precise score differences should be interpreted cautiously, the overall pattern of substantial and persistent performance gaps identified in Paper 1 remains a robust and worrying finding. These results underscore the urgent need for targeted interventions and policy reforms to address the underlying factors contributing to these disparities.

Rather than diminishing the relevance of the findings from this thesis, these issues further reinforce the overarching concerns with PIRLS data for low performing countries like South Africa. Indeed, the pervasive targeting issues identified in Paper 2, which revealed that a considerable proportion of the items were too difficult for the majority of South African students, were a central finding of this thesis. Importantly, the severity of these targeting issues – particularly for the lowest performing students – means that it is unlikely that a different methodological approach would substantially alter the fundamental finding that PIRLS is too difficult for the majority of South African students. Moreover, in light of the limitations of the data exposed in Paper 2, the findings from Paper 1 further expose the overarching concern that PIRLS data lacks the necessary nuance to appropriately reflect the unique educational landscape and challenges in South Africa. The concerns that result from the low overall performance, therefore, underscore the need for adapting or expanding the questionnaire items, as well as the international scaling techniques, to enhance the relevance and validity of the data for local contexts. This is particularly important for developing countries where reliability concerns from low performance have already been observed in previous cycles.

Another overarching limitation of this thesis is the decision to focus on three language groups – Afrikaans, English, and Sepedi – in Papers 2 and 3. Although this was a carefully considered choice, taking into account overall performance differences as well as representation of an understudied language, this selection limits the generalisability of findings to other language groups. However, this research sought to understand the consequences of widening participation, and in this regard, the findings for the lowest-performing group, Sepedi, remain important. Despite the restricted generalisability, therefore, the severe impact on validity, reliability, and comparability observed for the Sepedi group underscores the potential consequences for ensuring fair and equitable assessment of students in low-performing groups across all participating countries. This finding is especially relevant in contexts where PIRLS data is used to inform consequential decisions or justify targeted interventions.

Ultimately, although there are limitations regarding the analyses undertaken in this thesis, the overarching message regarding the importance of ensuring reliability, fairness, and validity for all students so that ILSA data can contribute nuanced, contextualised insights for participating countries, regardless of language or performance level, remains highly relevant.

9.2 Contributions

This research highlights the critical need for careful consideration of the challenges facing quality assessment, particularly in contexts where assessments may not adequately capture diverse ranges of student abilities. In exploring these challenges, the findings from this thesis make several important contributions to the field of educational assessment, particularly for evaluating the validity and comparability of ILSAs across linguistically diverse contexts. This section describes the key methodological contributions, the novel findings that emerged for the South African context, as well as the broader insights that can be gained from this investigation.

Firstly, although the use of multiple linear regression modelling to investigate the relationship between contextual factors and reading achievement is not novel, the contextualised approach to disentangling the complex multilingual educational landscape in South Africa provides a strong and comprehensive foundation for subsequent analyses of assessment results in the South African context. While the psychometric analyses in Papers 2 and 3 raised concerns about data robustness, the findings on language-based achievement gaps in Paper 1 remain significant, particularly its emphasis on the underperformance of African language groups compared to English and Afrikaans groups. Added to this, the consistency of these findings across cycles underscores the critical nature of these findings. While improving PIRLS targeting and contextualisation for South African students would enhance the quality and detail of future analyses, the current results already provide compelling and alarming evidence of systemic inequalities. These findings serve as a crucial indicator of educational disparities, emphasising the need for targeted interventions and policy reforms in the South African education system, particularly for students in African language groups.

The detailed contextual exploration also exposed some of the limitations of the PIRLS questionnaire data to adequately describe the intricate relationship between the factors influencing reading achievement in the South African context, which can help to inform areas where instruments could be adapted or improved to provide more relevant and valid data in the future. Added to this, Chapter 2 of this thesis provides a detailed description of the historical and current contextual considerations relevant for understanding the factors involved in reading achievement in the South African context. Together with this detailed contextual discussion, the analysis in Paper 1 contributes towards a more nuanced understanding of the factors influencing differential performance across language groups.

Secondly, the in-depth psychometric evaluation offers critical insights often overlooked in secondary analyses of ILSA data. This rigorous psychometric scrutiny has been identified as a key need for enhancing the validity and fairness of such assessments (Rutkowski & Rutkowski, 2021, 2022; Tijmstra et al., 2020; van Rijn et al., 2016). However, comprehensive evaluations of model fit, dimensionality, local independence, and student-item targeting across language groups for national subsets of data are rare – particularly for developing countries. This is a very relevant gap as developing countries often participate in ILSAs in order to gain insights into their own education system, and so ILSAs like PIRLS are frequently used to inform policy reforms or intervention efforts within countries (Addey & Sellar, 2019; Atar et al., 2023; Howie, 2022). The psychometric evaluation of fit undertaken in Paper 2, therefore, helps to provide insight into the limited validity of PIRLS in South Africa, as well as advising caution to policymakers and researchers when making inferences from the results.

Furthermore, the application of DRF methods allows for a novel examination of test comparability issues across languages within the South African context. While DIF analyses across language versions of PIRLS in South Africa have been undertaken for some language groups (e.g., Mtsatse & van Staden, 2021; Roux et al., 2022); extending this to evaluate the cumulative impact on passage and test-level comparability through DRF statistics is a distinct contribution. Passage-level comparability in particular is rarely examined for PIRLS, and investigations of differential functioning often select one passage to focus on without examining the potential impacts of differential functioning across passages. By focusing on passage-level comparability, therefore, this thesis exposed substantial concerns about the extent to which the passages designed to link the easier version of PIRLS to the main version functioned comparably across groups. This evaluative approach provides a holistic perspective on the impact of DIF issues that is often not considered in comparability studies; and had not yet been investigated in the South African context.

An added contribution regarding the language groups investigated is that there is a shortage of research regarding African languages generally, and Sepedi remains one of the least regularly investigated language groups in South Africa. Additionally, focusing on the differences across the highest and lowest performing language groups provides a more comprehensive understanding of the impact that low achievement can have on reliability, validity, and fairness across groups. The critical lens applied to the specific challenges associated with low overall achievement represents an important area of attention for ILSAs. By confronting the limitations introduced by these wide score distributions, this thesis presents a rigorous approach for understanding the impacts of low achievement on validity in other low-performing contexts.

Overall, the application of a contextualised analysis of the PIRLS results that directly addresses the unique challenges of the South African data, alongside a comprehensive psychometric evaluation of the instrument, provides a multi-faceted approach for evaluating the validity and comparability of ILSAs for diverse linguistic contexts.

9.3 Implications

Although the evaluation undertaken in this thesis focused specifically on PIRLS in South Africa, the findings have implications that extend beyond the South African context and speak to broader issues surrounding the use of ILSAs in linguistically diverse settings, particularly for low-performing education systems. This section explores some of the potential implications for ensuring quality assessment in ILSAs in the context of widening participation as well as recommendations for the future development and use of ILSAs.

9.3.1 Implications for international large-scale assessment development and design

The primary aims of ILSAs extend beyond just providing comparable information on selected skills and abilities across systems, many also seek to contribute towards an understanding of how those skills relate to broader social and economic outcomes (Kirsch et al., 2013). As a result of this there are extensive, and ever-evolving, practices of collecting the most relevant and useful background data to help ensure comprehensive evidence is provided. A challenge, however, is that there is an imbalance in the economic conditions of participating countries. Although the number of low and middle-income countries opting in to participate in ILSAs is growing, they still form the minority of participating countries. This means that decisions about criteria to include in international data collection tools such as the background questionnaires are more likely to address the needs of the majority, higher-income countries. In some cases, this challenge is further exacerbated by political associations. In PISA, for instance, major decisions about priority areas of focus or assessment adaptations are made by the governing board, however, only member OECD countries are eligible to vote on the governing board (Rutkowski & Rutkowski, 2021). Practices such as these automatically exclude needs specific to lower income countries, who are not eligible to join organisations like the OECD.

The challenges with the applicability of questionnaire items to the unique needs of different countries is not only a challenge for developing countries. Concerns about the relevance of ILSA data have been raised by a number of researchers looking at the applicability of background measures in higher-income settings in Europe, North America, and East Asia (Goldstein, 2017; Gomendio & Wert, 2023; Volante, 2017). Although the incorporation of national adaptations in contextual questionnaires to better suit local needs is already existing practice in the development of most ILSAs, these adaptations are typically then excluded or rescaled for international comparability in the public dataset (see Section 8.1.1). Added to this, the option to make changes to background questionnaires is used to different extents by participating countries, with some countries making multiple adaptations to the background questionnaires to best suit their needs, and others making little to no adaptations (see Foy, 2018b). There are also examples of countries opting out of collecting information via background questionnaires entirely. England, for example, chose not to distribute the home questionnaire for PIRLS 2021, opting instead to match the PIRLS results to background data collected through their own comprehensive National Pupil Database (Lindorff et al., 2023). By doing this, England's NRC is able to conduct detailed analyses with locally gathered data that is more attuned to their needs and interests. Questions remain, therefore, about the extent to which – even with the allowance for national adaptations on the background questionnaires – ILSA instruments are able to meet the diverse needs of participating countries.

Adaptive practices are not just related to the collection and integration of varied background information, they are also hailed as a means of capturing the wide ranges of ability (von Davier et al., 2020). In most cases, these practices rely on having item sets that represent different levels of difficulty and a number of linking items that overlap across difficulty levels which can then be used to establish a shared scale across very wide ranges of ability levels.

This is already an established practice in many ILSAs, particularly for participants at the lower end of the ability range. However, the issues identified across the linking passages in Paper 3 raise concerns about these adaptive testing practices more broadly. Similar to the problems with the differential levels of difficulty for shared items observed in Paper 3, researchers have also identified problems with shared items between PISA and PISA-D. Rutkowski and Rutkowski (2021) found extreme targeting issues related to PISA-D for countries with very low overall performance.

Additionally, in Spain, items used in PISA 2018 from PISA-D, which were designed to assess cognitive skills associated with reading fluency, were answered irregularly by students compared to their responses in later (more suitably targeted) sections of the reading assessment. The impact of these irregular response patterns meant that the reading results and data for Spain were not initially released for PISA 2018 (Gomendio, 2021). Crucially, as Gomendio and Wert (2023) note, since these problematic items were part of the adaptive design of PISA in that cycle, these anomalies could have additional consequences, not only leading to low scores but also potentially causing students to be misassigned to easier or more difficult items for the next part of the assessment.

Adaptive techniques, which rely on large numbers of well-functioning linking items to measure wide ability ranges validly and reliably, are increasingly being adopted by ILSA agencies and are regarded as the future of assessment (Frey et al., 2023; Rutkowski et al., 2022; von Davier et al., 2020; Yamamoto, Khorramdel, et al., 2018; Yamamoto, Shin, et al., 2018). It is therefore essential to consider the efficacy of these methods for all students, particularly those at the lower ends of the ability ranges, to ensure fair, valid, and reliable assessment practices are upheld.

9.3.2 Implications for the use of international large-scale assessment results

ILSAs are widely used by policymakers and education stakeholders around the world to inform decision-making and drive reforms at the national level (Arzarello et al., 2015; Fischman et al., 2019; Goldstein, 2017; Gomendio, 2021; Klieme, 2013; Lenkeit & Caro, 2014; Sjøberg, 2015; Spaull & Pretorius, 2022; Steiner-Khamsi & Waldow, 2018; Volante, 2017). Given their widespread influence, the implications of the findings from this thesis regarding the validity, reliability, and fair use of ILSA results, especially for low-performing countries and linguistically diverse contexts, are important to consider. As ILSAs continue to expand their global participation, ensuring the evidence they provide can be appropriately interpreted and translated into effective policies tailored to local contexts is essential.

One of the key areas of influence is the use of ILSA results as agents to inform and support equity-driven policy reform in education. Indeed, PIRLS has been promoted as a key tool that can be used in the realisation of SDG 4 – Quality education for all (UNESCO & IEA, 2017). In their description of the influential role PISA has played among European Union member states, Volante et al. (2022) highlights how findings from PISA have consistently led to an amplified focus on promoting educational equity – specifically targeting improvement efforts towards subpopulations performing at the lower end of the achievement range.

The substantial challenges identified with the precision of the estimation of achievement scores for low performing groups means that, without ensuring the validity and reliability of the scores within the relevant national context, ILSA results and subsequent reforms can inadvertently misguide or constrain plans that aim to promote equity. A potential challenge is that the lack of precision when estimating achievement scores for subgroups within countries could lead to the misallocation of resources based on ILSA scores. Added to this, the lack of comparability across groups could also lead to incorrect assumptions about ability gaps that do not consider the differing levels of difficulty faced by students in certain subgroups. The consequences of misguided reforms in developing countries – where educational infrastructure is often still in need of major investment and development – is of particular concern. If increased attention and budget is allocated to groups or areas based on the results of findings that are not reliable, the consequences may be longstanding and severe.

The potential for unsuccessful expenditure in education does not only relate to the possibility of misallocated resources due to imprecise or invalid measurements from ILSAs. These assessments also aim to provide insights into the relative strength and weaknesses of different education systems (Martin et al., 2019; OECD, 2023). However, some have argued that this intention to compare across countries has led to the practice of idolising the pedagogy and practices of ‘model’ countries – typically high-performing countries like Finland and Singapore – without paying adequate attention to contextual differences (Araújo et al., 2017; Waldow, 2009). Although it is challenging to establish a direct link between ILSA results and policy changes across contexts, there is evidence of comparisons, pedagogical change and policy borrowing from high performing countries (Auld & Morris, 2014; Brozo et al., 2014; Jude et al., 2021; Koh, 2018; Lawrence, 2020). For instance, there has been popular attention given to reducing class sizes as a means of improving educational outcomes; however, while some studies have shown that student-teacher ratios can influence student outcomes, these findings differ substantially across countries, grade levels, and schools (Duflo et al., 2015; Gazeaud & Ricard, 2024; Glewwe & Muralidharan, 2016; Gomendio & Wert, 2023; Justman, 2018; Vainikainen et al., 2017).

Some critics suggest that instead of being used to find what works for improving learning, results from ILSAs are often used selectively by governments to support pre-existing policies, aims or educational goals (Greger, 2012; Sjøberg, 2015). This use of ILSA results as a tool to legitimise or justify educational reforms for political means has been raised as a matter of concern by a number of researchers (Arzarello et al., 2015; Fischman et al., 2019; Sjøberg, 2015). A dominant theme that emerges in these critiques is that ILSAs are often used to provide evidence of causal relationships between pedagogy and student outcomes, when the cross-sectional data from the majority of ILSAs is ill-equipped to deliver causal conclusions about educational systems (Araújo et al., 2017; Eivers, 2010; Jerrim, 2011; Klieme, 2020).

This issue of drawing inappropriate causal inferences from ILSA data is particularly relevant to the widely researched topic of how language of instruction influences learning outcomes – a subject with major policy implications for countries that have a colonial history. This thesis identified two key language-related concerns from PIRLS results: the influence of test language on overall scores, and the lack of comparability across language versions. In particular, Paper 1 revealed that students who took the test in English achieved higher overall scores than those who were tested in an African language. This finding contrasts with extensive research across multiple countries (including South Africa) which indicates that students, especially in the early years of education, benefit from receiving instruction in their mother tongue rather than being transitioned too soon to English (e.g., Alkateb-Chami, 2024; Badenhorst & van der Merwe, 2017; Brock-Utne et al., 2019; Laitin et al., 2019; Opare-Kumi, 2024; Seid, 2019; Taylor & von Fintel, 2016).

Despite evidence showing the benefits of mother-tongue instruction and the risks of early transitions to English, there are still a number of countries – particularly in Africa – that use English as a medium of instruction even when it is not a common home language in the region (Clegg & Simpson, 2016; Coetzee-Van Rooy, 2018; James, 2021; Kretzer & Kaschula, 2021; Opare-Kumi, 2024). The widespread use of English as a medium of instruction raises concerns that findings from ILSAs showing higher test scores for students learning in English could be used to justify 'English-first' policies without considering the broader evidence on the benefits of mother-tongue instruction, or the concerns with the reliability, validity, and comparability of data from ILSAs for linguistically diverse contexts. Overlooking these issues could lead to problematic reforms or policies that disregard the potential advantages of mother-tongue instruction, particularly in the early years of education.

Another concerning implication of the widespread influence and use of ILSA results is the potential for these assessments to essentially define (or redefine) the priority areas for education in participating countries (Eivers, 2010; Sjøberg, 2015; Volante & Klinger, 2022). As Zhao (2020) argues, by globally promoting what PISA claims to measure as key skills and competencies, national education systems may no longer critically examine or debate the underlying goals of education in their local contexts. Rather, they narrowly focus on finding ways to improve performance on what the ILSA measures. This has been raised as a particular concern for developing countries, where shifting priorities to better align with the aims and ambitions of organisations such as the OECD may override the priorities of organisations such as UNESCO or UNICEF that have mandates specifically aimed at improving education and the lives of children in developing countries (Meyer & Zahedi, 2014; Sjøberg, 2015).

The implications associated with the potential for priorities and policies to change based on ILSA results are amplified by the environment of pressure and accountability that governments face to improve their ‘rankings’ (Breakspear, 2012; Jude et al., 2021; Klemenčič & Mirazchiyski, 2018). There have been multiple examples of the perception of underperformance in ILSAs being a powerful catalyst for policy reform. For instance, what is often described as “PISA shock” led to increased attention on educational priorities in Germany with specific focus on reducing achievement gaps and implementing national benchmarks (Greger, 2012; Sälzer & Prenzel, 2014). In Denmark, lower than expected achievement in PISA led to a shift in educational goals, with attention given to ensuring Denmark come ‘top 5’ in PISA rankings (Moos, 2014). While greater emphasis on enhancing educational policies and reducing achievement gaps is generally beneficial, this accountability dynamic has raised concerns about narrowing the focus of education policy simply to raise ILSA scores, rather than promoting more holistic improvement efforts (Gorur & Wu, 2015; van Staden et al., 2019; Zhao, 2020). For developing countries that are already grappling with systemic educational challenges, an emphasis on boosting rankings could result in counterproductive reforms that are detached from local needs, and potentially misaligned with the local curriculum expectations (Howie, 2012, 2022; Jude et al., 2021).

The findings from this thesis point toward several important areas that warrant further investigation to strengthen the validity and utility of ILSAs moving forward. Foremost is a need for extensive research focused on evaluating approaches to enhance the fairness of these assessments across diverse national contexts. This could include critically examining the existing translation and adaptation procedures to ensure they are appropriate for increasingly diverse linguistic environments. Additionally, more work is needed to establish comprehensive validity evidence and best practice guidelines specifically tailored to low-performing contexts, including appropriate methods for targeted measurement, scoring, and interpretation of results.

Research should also prioritise investigating how stakeholders like teachers, parents, and students in varied contexts perceive and experience ILSAs to better understand their local resonance and impacts. Although the relationship between ILSAs and policy has received considerable attention in developed countries, there is less attention and scrutiny given to this relationship in countries at the lower end of the performance scale. There would also be value in exploring potential models to promote collaborative partnerships between testing organisations and participating countries, particularly for countries that are not well represented in decision-making bodies like the OECD.

Test developers themselves have raised concern about the reliability of the results for low performers, even when adaptive testing procedures were used – as was demonstrated in the most recent PIRLS 2021 cycle for South Africa and Egypt (Mullis et al., 2023). The ethical implications of allowing or excluding such countries from participating in ILSAs altogether must therefore also be weighed. If participation continues, developing clear guidelines or best practices for interpreting and using ILSA results in low-performing contexts, becomes essential. It is not enough to acknowledge the potential limitations and state caveats in footnotes without providing clear guidance about what that means for the use of test results, particularly in contexts where the results hold high status for policymakers.

This thesis is by no means the first to critique the validity and fairness of evidence provided by ILSAs. Indeed, researchers have long called for increased reflection, accountability, and action related to the trust and reliance put on evidence from ILSAs (Araujo et al., 2017; Choi & Jerrim, 2016; Gomendio, 2023; Klemenčič & Mirazchiyski, 2018; Meyer & Zahedi, 2014; Rappleye et al., 2020; Volante, 2017; Zhao, 2016, 2020). However, the findings from this thesis underscore the extent to which the challenges raised regarding the evidence provided by ILSAs are even further exacerbated for low performing countries. As the range of participating countries diversifies, it is essential that ILSAs concentrate on evolving to ensure that no students or contexts are unfairly assessed, particularly as the current options for low performers are not sufficient. It is, however, not just the responsibility of testing organisations to ensure their assessments are fairly and validly used. Participating countries need to take care that they are using the results of ILSAs in valid and relevant ways to ensure that the choice to participate is both valuable and meaningful. Evidence is necessary for making effective, informed policy decisions for educational improvement. However, if those decisions are going to result in effective improvements, it is essential that they are based on valid, reliable, and fair evidence that is well suited to contextual needs.

REFERENCES

Addey, C., Maddox, B., & Zumbo, B. D. (2020). Assembled validity: Rethinking Kane's argument-based approach in the context of international large-scale assessments. *Assessment in Education: Principles, Policy & Practice*, *27*(6), 588–606. https://doi.org/10.1080/0969594X.2020.1843136

Addey, C., & Sellar, S. (2019). *Rationales for (non) participation in international large-scale learning assessments* (Education Research and Foresight: Working Paper No. 24). UNESCO.

Addey, C., Sellar, S., Steiner-Khamsi, G., Lingard, B., & Verger, A. (2017). The rise of international large-scale assessments and rationales for participation. *Compare*, *47*(3), 434–452. https://doi.org/10.1080/03057925.2017.1301399

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. American Educational Research Association, the American Psychological Association and the National Council on Measurement in Education.

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association, the American Psychological Association and the National Council on Measurement in Education.

Ahmed, S. K., Belisle, M., Cassity, E., Friedman, T., Lietz, P., & Spink, J. (2022). The role of international large-scale assessments (ILSAs) in economically developing countries. In T. Nilsen, A. Stancel-Piątak, & J.-E. Gustafsson (Eds.), *International handbook of comparative large-scale studies in education: Perspectives, methods and findings* (pp. 120–139). Springer.

Akkakoson, S. (2013). The relationship between strategic reading instruction, student learning of L2-based reading strategies and L2 reading achievement. *Journal of Research in Reading*, *36*(4), 422–450. https://doi.org/10.1111/jrir.12004

Alderson, J. C. (1985). Reading in a foreign language: A reading problem or a language problem. *RELC Journal*, *16*(2), 114–116.

Alkateb-Chami, M. (2024). Learning poverty when schools do not teach in children's home language: A comparative perspective. *International Journal of Educational Development*, *105*, 102975. https://doi.org/10.1016/J.IJEDUDEV.2023.102975

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement*, *36*(3), 185–198.

Almaskut, A., LaRoche, S., & Foy, P. (2023). Sample design in PIRLS 2021. In *Methods and procedures: PIRLS 2021 technical report* (pp. 1–32). Boston College, TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.tr2103.kb9560

Andersen, N. (2017, December 5). Education crisis: Eight out of ten South African children can't read – PIRLS report. *The South African*. https://www.thesouthafrican.com/news/eight-out-of-10-south-african-children-cant-read-pirls-report/

Andrich, D., & Marais, I. (2019). Violations of the assumption of independence I – Multidimensionality and response dependence. In *Rasch measurement theory: Measuring in the educational, social and health sciences* (pp. 173–185). https://doi.org/10.1007/978-981-13-7496-8_14

Andringa, S., & Godfroid, A. (2020). Sampling bias and the problem of generalizability in applied linguistics. *Annual Review of Applied Linguistics*, *40*, 134–142. https://doi.org/10.1017/S0267190520000033

APA, AERA, & NCMUE. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin*, *51*(2).

Araújo, L., & Costa, P. (2015). Home book reading and reading achievement in EU countries: the Progress in International Reading Literacy Study 2011 (PIRLS). *Educational Research and Evaluation*, *21*(5–6), 422–438. https://doi.org/10.1080/13803611.2015.1111803

Araújo, L., Morais, J., & Costa, A. (2013). Assessments of reading comprehension: Different measures, different explanatory variables and achievement. *L1-Educational Studies in Language and Literature*, *13*.

Araújo, L., Saltelli, A., & Schnepf, S. V. (2017). Do PISA data justify PISA-based education policy? *International Journal of Comparative Education and Development*, *19*(1), 20–34. https://doi.org/10.1108/IJCED-12-2016-0023

Ardington, C., & Henry, J. (2022). *Funda Wande Western Cape evaluation: Baseline report*. SALDRU, University of Cape Town.

Ardington, C., Wills, G., Pretorius, E. J., Deghaye, N., Mohohlwane, N. L., Menendez, A., Mtsatse, N., & van der Berg, S. (2020). *Technical Report: Benchmarking early grade reading skills in Nguni languages.* ReSEP, Stellenbosch University. SALDRU, University of Cape Town. NORC, University of Chicago. Department of Basic Education.

Ardington, C., Wills, G., Pretorius, E. J., Mohohlwane, N. L., & Menendez, A. (2021). Benchmarking oral reading fluency in the early grades in Nguni languages. *International Journal of Educational Development*, *84*, 102433. https://doi.org/10.1016/J.IJEDUDEV.2021.102433

Arffman, I. (2010). Equivalence of translations in international reading literacy studies. *Scandinavian Journal of Educational Research*, *54*(1), 37–59. https://doi.org/10.1080/00313830903488460

Arffman, I. (2013). Problems and issues in translating international educational achievement tests. *Educational Measurement: Issues and Practice*, *32*(2), 2–14.

Arzarello, F., Garuti, R., & Ricci, R. (2015). The impact of PISA studies on the Italian national assessment system. In K. Stacey & R. Turner (Eds.), *Assessing Mathematical Literacy* (pp. 249–260). Springer.

Atar, B., Atalay Kabasakal, K., & Kibrislioglu Uysal, N. (2023). Comparability of TIMSS 2015 mathematics test scores across country subgroups. *Journal of Experimental Education*, *91*(1), 82–100. https://doi.org/10.1080/00220973.2021.1913978

Aturupane, H., Glewwe, P., & Wisniewski, S. (2013). The impact of school quality, socioeconomic factors, and child health on students' academic performance: Evidence from Sri Lankan primary schools. *Education Economics*, *21*(1), 2–37. https://doi.org/10.1080/09645292.2010.511852

Aukerman, M., & Chambers Schuldt, L. (2021). What matters most? Toward a robust and socially just science of reading. *Reading Research Quarterly*, *56*(S1), S85–S103. https://doi.org/10.1002/RRQ.406

Auld, E., & Morris, P. (2014). Comparative education, the 'New Paradigm' and policy borrowing: constructing knowledge for educational reform. *Comparative Education*, *50*(2), 129–155. https://doi.org/10.1080/03050068.2013.826497

Badat, S., & Sayed, Y. (2014). Post-1994 South African education: The challenge of social justice. *ANNALS of the American Academy of Political and Social Science*, *652*(1), 127–148. https://doi.org/10.1177/0002716213511188

Badenhorst, J., & van der Merwe, R. (2017). Trial and error? An inquiry into implementing a 'straight-for-English' approach with foundation phase students with inadequate English proficiency. *Journal of Asian and African Studies*, *52*(7), 1040–1056. https://doi.org/10.1177/0021909616630994

Bailey, P., & Cohen, M. (2019). Statistical methods used in EdSurvey. In Lee, M., Zhang, T., Bailey, P., Buehler, E., Fink, T., Huo, H., Lee, S.-J., Liao, Y., & Sikali, E. (Eds.), *Analyzing NCES Data Using EdSurvey: A User's Guide*. American Institutes for Research.

Bailey, P., Emad, A., Huo, H., Lee, M., Liao, Y., Lishinski, A., Nguyen, T., Xie, Q., Yu, J., Zhang, T., Buehler, E., Lee, S., & Webb, B. (2023). Package 'EdSurvey': Analysis of NCES Education survey and assessment data. *R Package Version 4.0.4*.

Baird, J.-A., Isaacs, T., Johnson, S., Stobart, G., Yu, G., Sprague, T., & Daugherty, R. (2011). *Policy effects of PISA*. Oxford University Centre for Educational Assessment.

Bansilal, S. (2012). What can we learn from the KZN ANA results? *SA-EDUC Journal*, *9*(2).

Baroutsis, A., & Lingard, B. (2017). Counting and comparing school performance: An analysis of media coverage of PISA in Australia, 2000–2014. *Journal of Education Policy*, *32*(4), 432–449. https://doi.org/10.1080/02680939.2016.1252856

Barrett, A. M., & Tikly, L. (Eds.). (2013). *Education quality and social justice in the global South: Challenges for policy, practice and research*. Routledge. https://doi.org/10.4324/9780203817650

Bartram, D., & Hambleton, R. K. (2016). The ITC Guidelines: International standards and guidelines relating to tests and testing. In F. T. L. Leong, D. Bartram, F. Cheung, K. F. Geisinger, & D. Iliescu (Eds.), *The ITC international handbook of testing and assessment* (pp. 35–46). Oxford University Press. https://doi.org/10.1093/MED:PSYCH/9780199356942.003.0004

Bauer, D. J. (2023). Enhancing measurement validity in diverse populations: Modern approaches to evaluating differential item functioning. *British Journal of Mathematical and Statistical Psychology*, *76*(3), 435–461. https://doi.org/10.1111/BMSP.12316

Baumann, J. F. (2014). Vocabulary and reading comprehension: The nexus of meaning. In S. E. Israel & G. G. Duffy (Eds.), *Handbook of research on reading comprehension* (pp. 347–370). Routledge.

Baye, A., & Monseur, C. (2016). Gender differences in variability and extreme scores in an international context. *Large-Scale Assessments* in Education, 4(1), 1–16. https://doi.org/10.1186/s40536-015-0015-x

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/J.2517-6161.1995.TB02031.X

Berger, N. S., & Perfetti, C. A. (1977). Reading skill and memory for spoken and written discourse. *Journal of Literacy Research*, *9*(1), 7–16. https://doi.org/10.1080/10862967709547200

Bernhardt, E. B. (1993). Reading development in a second language: Theoretical, empirical & classroom perspectives. Ablex Publishing Corporation.

Bieleke, M., Goetz, T., Krannich, M., Roos, A. L., & Yanagida, T. (2023). Starting tests with easy versus difficult tasks: Effects on appraisals and emotions. *The Journal of Experimental Education*, *91*(2), 317–335. https://doi.org/10.1080/00220973.2021.1947764

Birnbaum, Z. W. (1969). *On the importance of different components in a multicomponent system* (Technical Report No. 54). Laboratory of Statistical Research, University of Washington.

Blair, T. R., Rupley, W. H., & Nichols, W. D. (2007). The effective teacher of reading: Considering the "what" and "how" of instruction. *The Reading Teacher*, *60*(5), 432–438. https://doi.org/10.1598/RT.60.5.3

Boakye, N. (2015). The social dimension of reading literacy development in South Africa: Bridging inequalities among the various language groups. *International Journal of the Sociology of Language*, *2015*(234), 133–156. https://doi.org/10.1515/ijsl-2015-0008

Bond, P. (2003). The comparative apologetics of racially regressive laws in the Confederate States of America and the South African Republic or How did they live with themselves. *Journal of Law and Social Change*, *7,* 29-49.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press. https://doi.org/10.1017/CBO9780511490026

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791–799. https://doi.org/10.1080/01621459.1976.10480949

Breakspear, S. (2012). *The policy impact of PISA: An exploration of the normative effects of international benchmarking in school system performance* (OECD Education Working Papers No. 71). https://doi.org/10.1787/5k9fdfqffr28-en

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*, *38*(4), 295–317. https://doi.org/10.1111/j.1745-3984.2001.tb01129.x

Bristol, J., Mullis, I. V. S., Fishbein, B., & Foy, P. (2023). Reviewing the PIRLS 2021 achievement item statistics. In M. von Davier, I. V. S. Mullis, B. Fishbein, & P. Foy (Eds.), *Methods and procedures: PIRLS 2021 technical report* (pp. 1–45). Boston College, TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.tr2103.kb5892

Brochu, P., O'Grady, K., Scerbina, T., & Tao, Y. (2017). PIRLS/ePIRLS 2016 Canada in context: Canadian results from the Progress in International Reading Literacy Study. Council of Ministers of Education, Canada.

Brock-Utne, B., Desai, Z., Qorro, M. A. S., & Pitman, A. (Eds.). (2019). *Language of Instruction in Tanzania and South Africa – Highlight from a Project*. Sense Publishers. https://doi.org/10.1163/9789460912221

Broom, Y. (2004). Reading English in multilingual South African primary schools. *International Journal of Bilingual Education and Bilingualism*, *7*(6), 506–528. https://doi.org/10.1080/13670050408667828

Brozo, W. G., Sulkunen, S., Shiel, G., Garbe, C., Pandian, A., & Valtin, R. (2014). Reading, gender, and engagement: Lessons from five PISA countries. *Journal of Adolescent & Adult Literacy*, *57*(7), 584–593.

Buckingham, J., Beaman, R., & Wheldall, K. (2014). Why poor children are more likely to become poor readers: The early years. *Educational Review*, *66*(4), 428–446. https://doi.org/10.1080/00131911.2013.795129

Bulut, O., & Suh, Y. (2017). Detecting multidimensional differential item functioning with the multiple indicators multiple causes model, the item response theory likelihood ratio test, and logistic regression. *Frontiers in Education*. https://doi.org/10.3389/feduc.2017.00051

Burgess, S. R., Hecht, S. A., & Lonigan, C. J. (2002). Relations of the home literacy environment (HLE) to the development of reading-related abilities: A one-year longitudinal study. *Reading Research Quarterly*, *37*(4), 408–426. https://doi.org/10.1598/RRQ.37.4.4

Burton, R. F. (2001). Quantifying the effects of chance in multiple choice and true/false tests: Question selection and guessing of answers. *Assessment & Evaluation in Higher Education*, *26*(1), 41–50. https://doi.org/10.1080/02602930020022273

Cain, K., & Oakhill, J. (2006). Assessment matters: Issues in the measurement of reading comprehension. *British Journal of Educational Psychology*, *76*(4), 697–708. https://doi.org/10.1348/000709905X69807

Camilli, G. (2013). Ongoing issues in test fairness. *Educational Research and Evaluation*, *19*(2–3), 104–120.

Casale, D., & Posel, D. (2011). English language proficiency and earnings in a developing country: The case of South Africa. *The Journal of Socio-Economics*, *40*(4), 385–393. https://doi.org/10.1016/J.SOCEC.2011.04.009

Castles, A., Rastle, K., & Nation, K. (2018). Ending the Reading Wars: Reading acquisition from novice to expert. *Psychological Science in the Public Interest*, *19*(1), 5–51. https://doi.org/10.1177/1529100618772271

Cele, N. (2001). Oppressing the oppressed through language liberation: Repositioning English for meaningful education. *Perspectives in Education*, *19*(1), 181–194. https://journals.co.za/content/persed/19/1/EJC87051

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*, 1–29.

Chalmers, R. P. (2018). Model-based measures for detecting and quantifying response bias. *Psychometrika*, *83*(3), 696–732. https://doi.org/10.1007/s11336-018-9626-9

Chalmers, R. P. (2022). A unified comparison of IRT-based effect sizes for DIF investigations. *Journal of Educational Measurement, 60*(2), 318-350. https://doi.org/10.1111/jedm.12347

Chalmers, R. P., Counsell, A., & Flora, D. B. (2016). It might not make a big DIF: Improved differential test functioning statistics that account for sampling variability. *Educational and Psychological Measurement*, *76*(1), 114–140. https://doi.org/10.1177/0013164415584576

Chansa-Kabali, T., Serpell, R., & Lyytinen, H. (2014). Contextual analysis of home environment factors influencing the acquisition of early reading skills in Zambian families. *Journal of Psychology in Africa*, *24*(5), 410–419. https://doi.org/10.1080/14330237.2014.997008

Chapman, J. W., Tunmer, W. E., & Prochnow, J. E. (2000). Early reading-related skills and performance, reading self-concept, and the development of academic self-concept: A longitudinal study. *Journal of Educational Psychology*, *92*(4), 703–708. https://doi.org/10.1037/0022-0663.92.4.703

Chen, C., & Hwu, B. S. (2018). Improving the assessment of differential item functioning in large-scale programs with dual-scale purification of Rasch models: The PISA example. *Applied Psychological Measurement*, *42*(3), 206–220. https://doi.org/10.1177/0146621617726786

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289. https://doi.org/10.3102/10769986022003265

Chilenga-Butao, T., Pakade, N., & Ehren, M. (2020). National standardized assessments in South Africa: Policy and power play. In S. Grek, C. Maroy, & A. Verger (Eds.), *World yearbook of education 2021: Accountability and datafication in the governance of education* (pp. 125–142). Routledge.

Chinapah, V., H'ddigui, E. M., Kanjee, A., Falayajo, W., Fomba, C. O., Hamissou, O., Rafalimanana, A., & Byomugisha, A. (2000). With Africa for Africa: Towards quality education for all. *1999 MLA Project*.

Chisholm, L., & Wildeman, R. (2013). The politics of testing in South Africa. *Journal of Curriculum Studies*, *45*(1), 89–100. https://doi.org/10.1080/00220272.2012.755710

Chiu, M. M., & Chow, B. W. Y. (2010). Culture, motivation, and reading achievement: High school students in 41 countries. *Learning and Individual Differences*, *20*(6), 579–592. https://doi.org/10.1016/j.lindif.2010.03.007

Chmielewski, A. K. (2019). The global increase in the socioeconomic achievement gap, 1964 to 2015. *American Sociological Review*, *84*(3), 517–544. https://doi.org/10.1177/0003122419847165

Choi, Á., & Jerrim, J. (2016). The use (and misuse) of PISA in guiding policy reform: the case of Spain. *Comparative Education*, *52*(2), 230–245. https://doi.org/10.1080/03050068.2016.1142739

Christianson, K., & Luke, S. G. (2011). Context strengthens initial misinterpretations of text. *Scientific Studies of Reading*, *15*(2), 136–166.

Clarke, P. J., Snowling, M. J., Truelove, E., & Hulme, C. (2010). Ameliorating children's reading-comprehension difficulties: A randomized controlled trial. *Psychological Science*, *21*(8), 1106–1116. https://doi.org/10.1177/0956797610375449

Clauser, B. E., & Bunch, M. B. (Eds.). (2021). *The history of educational measurement: Key advancements in theory, policy, and practice*. Routledge. https://doi.org/10.4324/9780367815318

Clegg, J., & Simpson, J. (2016). Improving the effectiveness of English as a medium of instruction in sub-Saharan Africa. *Comparative Education*, *52*(3), 359–374. https://doi.org/10.1080/03050068.2016.1185268

Coady, M. R., Makalela, L., & Lopez, M. P. S. (2022). Metaliteracy and writing among 4th grade multilingual students in South Africa. *International Journal of Multilingualism*, *19*(3), 435–455. https://doi.org/10.1080/14790718.2019.1631829

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), 19–46. https://doi.org/10.1080/10228195.2018.1493613

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, *42*(2), 133–148. https://doi.org/10.1111/j.1745-3984.2005.00007

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Routledge. https://doi.org/10.4324/9780203774441

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 201–219). American Council on Education.

Combrinck, C. (2020). Big changes in achievement between cohorts: A true reflection of educational improvement or is the test to blame? In M. S. Khine (Eds.) *Rasch measurement: Applications in quantitative educational research* (pp. 179–193). Springer. https://doi.org/10.1007/978-981-15-1800-3_10

Combrinck, C., & Mtsatse, N. (2019). Reading on paper or reading digitally? Reflections and implications of ePIRLS 2016 in South Africa. *South African Journal of Education*, *39*(2), S1-S14. https://doi.org/10.15700/saje.v39ns2a1771

Compton-Lilly, C., Spence, L. K., Thomas, P. L., & Decker, S. L. (2023). Stories grounded in decades of research: What we truly know about the teaching of reading. *Reading Teacher, 77*(3), 392-400. https://doi.org/10.1002/trtr.2258

Constitution of the Republic of South Africa, No 108 (1996).

Constitution Review Committee. (2017). Status of Sepedi, Sesotho sa Leboa, Khelovhedu & Sign Language of South Africa & Khoi, Nama & San Languages: Pansalb, DAC & CRL Rights Commission. Constitutional Review Committee Meeting. https://pmg.org.za/committee-meeting/23912/

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *The American Journal of Medicine*, *119*(2), 166.e7-166.e16. https://doi.org/10.1016/j.amjmed.2005.10.036

Cordero, J. M., & Mateos-Romero, L. (2021). Exploring factors related with resilience in primary education: Evidence from European countries. *Studies in Educational Evaluation*, *70*, 101045. https://doi.org/10.1016/j.stueduc.2021.101045

Costa, P., & Araújo, L. (2018). Skilled students and effective schools: Reading achievement in Denmark, Sweden, and France. *Scandinavian Journal of Educational Research*, *62*(6), 850–864. https://doi.org/10.1080/00313831.2017.1307274

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation*, *10*(1), 7. https://doi.org/10.7275/jyj1-4868

Crato, N. (Ed.). (2021). Improving a country's education: PISA 2018 results in 10 countries. Springer.

Cresswell, J., Schwantner, U., & Waters, C. (2015). *A review of international large-scale assessments in education: Assessing component skills and collecting contextual data* (Report). PISA, The World Bank, OECD Publishing. https://doi.org/10.1787/9789264248373-en

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Cronbach, L. J. (1971). Test validation. In E. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). American Council on Education.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281.

Crooks, T. J., Kane, M. T., & Cohen, A. S. (2010). Threats to the valid use of assessments. *Assessment in Education: Principles, Policy & Practice*, *21*(1), 265–286. https://doi.org/10.1080/0969594960030302

Crowe, E. C., Connor, C. M., & Petscher, Y. (2009). Examining the core: Relations among reading curricula, poverty, and first through third grade reading achievement. *Journal of School Psychology*, *47*(3), 187–214. https://doi.org/10.1016/J.JSP.2009.02.002

Cummins, J. (2012). The intersection of cognitive and sociocultural factors in the development of reading comprehension among immigrant students. *Reading and Writing*, *25*(8), 1973–1990. https://doi.org/10.1007/s11145-010-9290-7

Dai, S., Vo, T. T., Kehinde, O. J., He, H., Xue, Y., Demir, C., & Wang, X. (2021). Performance of polytomous IRT models with rating scale data: An investigation over sample size, instrument length, and missing data. *Frontiers in Education*, *6*, 721963. https://doi.org/10.3389/feduc.2021.721963

DBE. (2010). Education for All: 2009 Country report South Africa. Department of Basic Education.

DBE. (2011a). Curriculum and assessment policy statement Grades 1-3: English home language (Foundation phase). Department of Basic Education.

DBE. (2011b). Curriculum and assessment policy statement Grades 4-6: English home language. (Intermediate phase). Department of Basic Education.

DBE. (2011c). Curriculum and assessment policy statement Grades 7-9: English home language (Senior phase). Department of Basic Education.

DBE. (2013). The incremental introduction of African languages in South African schools. Department of Basic Education.

DBE. (2015). Action Plan to 2019: Towards the realisation of schooling 2030. *Taking forward South Africa's National Development Plan 2030*. Department of Basic Education.

DBE. (2017a, December 5). *Department of Basic Education response to release of Progress in International Reading Study 2016 results* [Press Release]. https://www.education.gov.za/Newsroom/MediaReleases/tabid/347/ctl/Details/mid/5986/ItemID/5522/Default.aspx

DBE. (2017b). *The development of a systemic evaluation model for the basic education sector*. Department of Basic Education.

DBE. (2021). Early Grade Reading Programme evaluation: Midline report 2021. Department of Basic Education.

DBE. (2022). *Synthesis report of EGRS I and EGRS II*. Department of Basic Education.

DBE. (2023). *PIRLS 2021: South Africa main report*. Department of Basic Education.

De Ayala, R. J., Plake, B. S., & Impara, J. C. (2001). The impact of omitted responses on the accuracy of ability estimation in item response theory. *Journal of Educational Measurement*, *38*(3), 213–234. https://doi.org/10.1111/J.1745-3984.2001.TB01124.X

De Backer, F., Slembrouck, S., & van Avermaet, P. (2019). Assessment accommodations for multilingual learners: pupils' perceptions of fairness. *Journal of Multilingual and Multicultural Development*, *40*(9), 833–846. https://doi.org/10.1080/01434632.2019.1571596

De Boeck, P., & Elosua Oliden, P. (2016). Reliability and validity: History, Notions, Methods, and Discussion. In F. T. L. Leong, D. Bartram, F. Cheung, K. F. Geisinger, & D. Iliescu (Eds.), *The ITC international handbook of testing and assessment* (pp. 408–422). Oxford University Press. https://doi.org/10.1093/med:psych/9780199356942.003.0028

DeMars, C. E. (2011). An analytic comparison of effect sizes for differential item functioning. *Applied Measurement in Education*, *24*(3), 189–209. https://doi.org/10.1080/08957347.2011.580255

DeMars, C. E., Bashkov, B. M., & Socha, A. B. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Research & Practice in Assessment*, *8*, 69–82.

DeMars, C. E., & Jurich, D. P. (2015). The interaction of ability differences and guessing when modeling differential item functioning with the Rasch model: Conventional and tailored calibration. *Educational and Psychological Measurement*, *75*(4), 610–633. https://doi.org/10.1177/0013164414554082

DeMars, C. E., & Wise, S. L. (2010). Can differential rapid-guessing behavior lead to differential item functioning? *International Journal of Testing*, *10*(3), 207–229. https://doi.org/10.1080/15305058.2010.496347

Deng, J., & Rios, J. A. (2022). Investigating the effect of differential rapid guessing on population invariance in equating. *Applied Psychological Measurement, 46*(7), 589-684. https://doi.org/10.1177/01466216221108991

Department of Education. (2003). Systemic evaluation: Foundation phase. Quality Assurance Department of Education.

Department of Education. (2008). Education for All: Country report South Africa. Department of Education, UNESCO.

Didion, L., Toste, J. R., & Filderman, M. J. (2020). Teacher professional development and student reading achievement: A meta-analytic review of the effects. *Journal of Research on Educational Effectiveness*, *13*(1), 29–66. https://doi.org/10.1080/19345747.2019.1670884

Dixon, L. Q., Wu, S., & Daraghmeh, A. (2012). Profiles in bilingualism: Factors influencing kindergartners' language proficiency. *Early Childhood Education Journal*, *40*(1), 25–34.

Dooley, C. M. (2017). The emergence of comprehension: A decade of research 2000-2010. *International Electronic Journal of Elementary Education*, *4*(1), 169–184. https://www.iejee.com/index.php/IEJEE/article/view/219

Dorans, N. J., & Cook, L. L. (Eds.). (2016). *Fairness in educational assessment and measurement* (1st ed.). Routledge. https://doi.org/10.4324/9781315774527

Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education*, *37*(9), 830–837. https://doi.org/10.1046/J.1365-2923.2003.01594.X

Duflo, E., Dupas, P., & Kremer, M. (2015). School governance, teacher incentives, and pupil-teacher ratios: Experimental evidence from Kenyan primary schools. *Journal of Public Economics*, *123*, 92–110. https://doi.org/10.1016/j.jpubeco.2014.11.008

Duke, N. K., & Carlisle, J. (2011). The development of comprehension. In M. L. Kamil, D. P. Pearson, E. B. Moje, & P. P. Afflerbach (Eds.), *Handbook of reading research: Vol. IV* (pp. 199–228). Routledge.

Dunn, L. M., & Dunn, D. M. (2007). *PPVT-4: Peabody picture vocabulary test* (4th ed.). Pearson Assessments.

Dunnette, M. D. (1963). A modified model for test validation and selection research. *Journal of Applied Psychology*, *47*(5), 317–323. https://doi.org/10.1037/H0047635

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 in PIRLS 2016*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Ebel, R. L. (1972). *Essentials of educational measurement* (2nd ed.). Prentice-Hall.

Ehren, M. (2021). International large-scale assessments and education system reform: On the power of numbers. In T. Nilsen, A. Stancel-Piątak, & J.-E. Gustafsson (Eds.), *International Handbook of Comparative Large-Scale Studies in Education* (pp. 97–118). Springer. https://doi.org/10.1007/978-3-030-38298-8_36-1

Eivers, E. (2010). PISA: Issues in implementation and interpretation. *The Irish Journal of Education / Iris Eireannach an Oideachais*, *38*(xxxviii), 94–118.

Eklöf, H. (2010). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy and Practice*, *17*(4), 345–356. https://doi.org/10.1080/0969594X.2010.516569

El Masri, Y., & Andrich, D. (2020). The trade-off between model fit, invariance, and validity: The case of PISA science assessments. *Applied Measurement in Education*, *33*(2), 174–188. https://doi.org/10.1080/08957347.2020.1732384

El Masri, Y., Baird, J.-A., & Graesser, A. C. (2016). Language effects in international testing: The case of PISA 2006 science items. *Assessment in Education: Principles, Policy & Practice*, *23*(4), 427–455. https://doi.org/10.1080/0969594X.2016.1218323

Elosua Oliden, P., & Mujika Lizaso, J. (2013). Invariance levels across language versions of the PISA 2009 reading comprehension tests in Spain. *Psicothema*, *25*(3), 390–395. https://doi.org/10.7334/psicothema2013.46

Embretson, S. E. (2007). Construct validity: A universal validity system or just another test evaluation procedure? *Educational Researcher*, *36*(8), 449–455. https://doi.org/10.3102/0013189X07311600

Embretson, S. E., & Reise, S. P. (2013). *Item response theory for psychologists*. Psychology Press. https://doi.org/10.4324/9781410605269

Engel, L. C., & Rutkowski, D. (2021). Costs of big data: Challenges and possibilities of cost-benefit analysis of ILSAs. In C. Wyatt-Smith, B. Lingard, & E. Heck (Eds.), *Digital disruption in teaching and testing* (pp. 124–135). Routledge. https://doi.org/10.4324/9781003045793-7

Ercikan, K. (1998). Translation effects in international assessments. *International Journal of Educational Research*, *29*(6), 543–553. https://doi.org/10.1016/S0883-0355(98)00047-0

Ercikan, K., & Allalouf, A. (2016). Assessment design for accuracy of scores, meaningfulness of interpretations, and fairness of decision-making in high-stakes educational testing. In F. T. L. Leong, D. Bartram, F. Cheung, K. F. Geisinger, & D. Iliescu (Eds.), *The ITC International Handbook of Testing and Assessment* (pp. 449–460). International Test Commission.

Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon, F., & Lacroix, S. (2010). Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews. *Educational Measurement: Issues and Practice*, *29*(2), 24–35. https://doi.org/10.1111/J.1745-3992.2010.00173.X

Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing*, *5*(1), 23–35.

Ercikan, K., & Por, H.-H. (2020). Comparability in multilingual and multicultural assessment contexts. In A. I. Berman, E. H. Haertel, & J. W. Pellegrino (Eds.), *Comparability of large-scale educational assessments: Issues and recommendations* (pp. 205–226). National Academy of Education. https://doi.org/10.31094/2020/1

Eriksson, K. (2014). Does the language of instruction in primary school affect later labour market outcomes? Evidence from South Africa. *Economic History of Developing Regions*, *29*(2), 311–335. https://doi.org/10.1080/20780389.2014.955272

Evans, R., & Nthulana, I. (2018). Linguistic challenges faced by rural Tshivenda-speaking teachers when Grade 4 learners transition to English. *The Journal for Transdisciplinary Research in Southern Africa*, *14*(2). https://doi.org/10.4102/TD.V14I2.545

Evans, V. (2011). Language and cognition: The view from cognitive linguistics. In V. Cook & B. Bassetti (Eds.), *Language and bilingual cognition* (pp. 83–122). Psychology Press.

Everatt, J., Smythe, I., Ocampo, D., & Gyarmathy, E. (2004). Issues in the assessment of literacy-related difficulties across language backgrounds: A cross-linguistic comparison. *Journal of Research in Reading*, *27*(2), 141–151. https://doi.org/10.1111/j.1467-9817.2004.00222.x

Faulkner-Bond, M., & Sireci, S. G. (2015). Validity issues in assessing linguistic minorities. *International Journal of Testing*, *15*(2), 114–135. https://doi.org/10.1080/15305058.2014.974763

Finch, W. H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement*, *45*(3), 225–245. https://doi.org/10.1111/J.1745-3984.2008.00062.X

Finch, W. H., & French, B. F. (2023). Effect sizes for estimating differential item functioning influence at the test level. *Psych*, *5*(1), 133–147. https://doi.org/10.3390/psych5010013

Finch, W. H., & Hernández Finch, M. E. (2016). Recursive partitioning to identify potential causes of differential item functioning in cross-national data. *International Journal of Testing*, *16*(1), 21–53. https://doi.org/10.1080/15305058.2015.1039644

Fischman, G. E., Topper, A. M., Silova, I., Goebel, J., & Holloway, J. L. (2019). Examining the influence of international large-scale assessments on national education policies. *Journal of Education Policy*, *34*(4), 470–499. https://doi.org/10.1080/02680939.2018.1460493

Fiske, E. B., & Ladd, H. F. (2004). *Elusive equity: Education reform in post-apartheid South Africa*. Brookings Institution Press.

Fleisch, B. (2008). Primary education in crisis: Why South African schoolchildren underachieve in reading and mathematics. Juta and Company.

Foy, P. (2018a). *PIRLS 2016 user guide for the international database*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Foy, P. (2018b). Supplement 2 national adaptations of the international context questionnaires. In P. Foy (2018a), *PIRLS 2016 user guide for the international database*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Foy, P., Martin, M. O., & Mullis, I. V. S. (2010, July 1-3). The limits of measurement: Problems in estimating reading achievement in PIRLS 2006 for low-performing countries [Conference paper]. The 4th IEA International Research Conference, Gothenburg, Sweden.

Foy, P., Martin, M. O., Mullis, I. V. S., & Yin, L. (2018). Reviewing the PIRLS 2016 achievement item statistics. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in PIRLS 2016*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Foy, P., & Yin, L. (2017). Scaling the PIRLS 2016 achievement data. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in PIRLS 2016*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Frey, A., König, C., & Fink, A. (2023). A highly adaptive testing design for PISA. *Journal of Educational Measurement*. https://doi.org/10.1111/jedm.12382

Fulmer, S. M., & Tulis, M. (2013). Changes in interest and affect during a difficult reading task: Relationships with perceived difficulty and reading fluency. *Learning and Instruction*, *27*, 11–20. https://doi.org/10.1016/j.learninstruc.2013.02.001

Gazeaud, J., & Ricard, C. (2024). Learning effects of conditional cash transfers: The role of class size and composition. *Journal of Development Economics*, *166*, 103194. https://doi.org/10.1016/J.JDEVECO.2023.103194

Gee, J. P. (1989). What is literacy? *Journal of Education*, *171*(1), 18–25.

Geva, E., Xi, Y., Massey-Garrison, A., & Mak, J. Y. (2019). Assessing reading in second language learners: Development, validity, and educational considerations. In D. Kilpatrick, R. Joshi, & R. Wagner (Eds.), *Reading development and difficulties* (pp. 117–155). Springer.

Glewwe, P., & Muralidharan, K. (2016). Improving education outcomes in developing countries: Evidence, knowledge gaps, and policy implications. In E. A. Hanushek, S. Machin, L. Woessmann (Eds.), *Handbook of the Economics of Education*: Vol 5 (pp. 653–743). https://doi.org/10.1016/B978-0-444-63459-7.00010-5

Gnambs, T., & Lenhard, W. (2023). Remote Testing of Reading Comprehension in 8-Year-Old Children: Mode and Setting Effects. *Assessment, 31*(2), 219-537. https://doi.org/10.1177/10731911231159369

Gökçe, S., Berberoğlu, G., Wells, C. S., & Sireci, S. G. (2021). Linguistic distance and translation differential item functioning on Trends in International Mathematics and Science Study Mathematics assessment items. *Journal of Psychoeducational Assessment*, *39*(6), 728–745. https://doi.org/10.1177/07342829211010537

Goldenberg, C. (2020). Reading wars, reading science, and English learners. *Reading Research Quarterly*, *55*(S1), S131–S144. https://doi.org/10.1002/RRQ.340

Goldstein, H. (2017). Measurement and evaluation issues with PISA. In L. Volante (Ed.), *The PISA Effect on Global Educational Governance*. Routledge.

Gomendio, M. (2021). The evidence provided by international large-scale assessments about the Spanish education system: Why nobody listens despite all the noise. In N. Crato (Ed.), *Improving a country's education: PISA 2018 results in 10 countries* (pp. 175–200). Springer.

Gomendio, M. (2023). PISA: Mission failure – With so much evidence from student testing, why do education systems continue to struggle? *Education Next*, *23*(2), 16–22.

Gomendio, M., & Wert, J. I. (2023). *Dire Straits: Education Reforms, Ideology, Vested Interests, and Evidence*. Open Book Publishers. https://doi.org/10.11647/OBP.0332

Goodrich, J. M., Lonigan, C. J., & Farver, J. A. M. (2013). Do early literacy skills in children's first language promote development of skills in their second language? An experimental evaluation of transfer. *Journal of Educational Psychology*, *105*(2), 414–426. https://doi.org/10.1037/A0031780

Goodrich, S., & Ercikan, K. (2019). Measurement comparability of reading in the English and French-Canadian populations: Special case of the 2011 Progress in International Reading Literacy Study. *Frontiers in Education*, *4*, 120. https://doi.org/10.3389/feduc.2019.00120

Goriot, C., van Hout, R., Broersma, M., Lobo, V., McQueen, J. M., & Unsworth, S. (2021). Using the Peabody Picture Vocabulary Test in L2 children and adolescents: Effects of L1. *International Journal of Bilingual Education and Bilingualism*, *24*(4), 546–568. https://doi.org/10.1080/13670050.2018.1494131

Gorsky, G. S. (2019). *Reading between the lines: Revealing and resisting the "hidden" gender bias in math story problems* [Doctoral dissertation, University of Washington]. ResearchWorks Archive. https://digital.lib.washington.edu/researchworks/handle/1773/44788

Gorur, R., & Wu, M. (2015). Leaning too far? PISA, policy and Australia's 'top five' ambitions. *Discourse: Studies in the Cultural Politics of Education*, *36*(5), 647–664. https://doi.org/10.1080/01596306.2014.930020

Gough, P. B. (1972). One second of reading. In J. F. Kavanagh & I. G. Mattingly (Eds.), *Language by ear and by eye: The relationship between speech and reading*. MIT Press.

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education, 7*(1), 4-61. https://doi.org/10.1177/074193258600700104

Gove, A., & Wetterberg, A. (Eds.). (2011). *The early grade reading assessment: An introduction*. RTI International.

Govender, P. (2023, May 16). South African children come last in international reading assessment. *Times Live*. https://www.timeslive.co.za/news/south-africa/2023-05-16-south-african-children-come-last-in-international-reading-assessment/

Govender, R., & Hugo, A. J. (2020). An analysis of the results of literacy assessments conducted in South African primary schools. *South African Journal of Childhood Education*, *10*(1), 1–13. https://doi.org/10.4102/sajce.v10i1.745

Greger, D. (2012). When PISA does not matter? The case of the Czech Republic and Germany. *Human Affairs*, *22*(1), 31–42. https://doi.org/10.2478/s13374-012-0004-5

Gregory, K., & Kanjee, A. (2009, September 9-13). *Modelling science achievement in South Africa using international achievement data: some comparability and validity issues* [Conference presentation]. 35th Annual Meeting of International Association for Educational Assessment, Brisbane, Australia.

Grisay, A., Gonzalez, E., & Monseur, C. (2009). Equivalence of item difficulties across national versions of the PIRLS and PISA reading assessments. *IERI: Issues and Methodologies in Large-Scale Assessments*, *2*, 63–83.

Grisay, A., Hal, J., Jong, D., Gebhardt, E., & Berezner, A. (2007). Translation equivalence across PISA countries. *Journal of Applied Measurement, 8*(3), 249-267.

Grömping, U. (2012). Variable importance assessment in regression: Linear regression versus random forest. *The American Statistician*, *63*(4), 308–319. https://doi.org/10.1198/TAST.2009.08199

Guion, R. M. (1977). Content validity – The source of my discontent. *Applied Psychological Measurement*, *1*(1), 1–10. https://doi.org/10.1177/014662167700100103

Guion, R. M. (1980). On Trinitarian doctrines of validity. *Professional Psychology: Research and Practice*, *11*(3), 385–398. https://doi.org/10.1037/0735-7028.11.3.385

Guo, H., & Ercikan, K. (2020). Differential rapid responding across language and cultural groups. *Educational Research and Evaluation*, *26*(5-6), 302–327. https://doi.org/10.1080/13803611.2021.1963941

Gustafsson, M. (2018). An overview of education policy 2007-2017.

Gustafsson, M. (2019a). Floor effects and the comparability of developing country student test scores (Working paper). ReSEP, Stellenbosch University.

Gustafsson, M. (2019b). Pursuing equity through policy in the schooling sector 2007–2017. In N. Spaull & J. Jansen (Eds.), *South African schooling: The enigma of inequality* (pp. 47–66). Springer. https://doi.org/10.1007/978-3-030-18811-5_3

Gustafsson, M. (2020). A revised PIRLS 2011 to 2016 trend for South Africa and the importance of analysing the underlying microdata (Working Paper No. 02). ReSEP, Stellenbosch University.

Gustafsson, M., & Barakat, B. F. (2023). Marginalized learners in international and regional test data: The extent of floor effects. *Comparative Education Review*, *67*(3). https://doi.org/10.1086/725442

Gustafsson, M., & Crouch, L. (2018). *Worldwide inequality and poverty in cognitive results: Cross-sectional evidence and time-based trends* (RISE Working Paper No. 18/019). Research on Improving Systems of Education.

Gustafsson, M., & Taylor, N. (2022). *The politics of improving learning outcomes in South Africa* (RISE Working Paper No. PE03). Research on Improving Systems of Education.

Gustafsson, M., & Taylor, S. (2023). Explaining improved learning outcomes in South African primary schools. In S. van Staden & C. Combrinck (Eds.), *Tracking changes in South African reading literacy achievement* (pp. 36–65). Brill.

Guzmán-Simón, F., Gil-Flores, J., & Pacheco-Costa, A. (2020). Home literacy environment and reading comprehension in Spanish primary education. *Journal of Research in Reading*, *43*(2), 229–247. https://doi.org/10.1111/1467-9817.12299

Haertel, E. H. (2013). Reliability and validity of inferences about teachers based on student scores: William H. Angoff memorial lecture series. *Educational Testing Service*.

Haertel, E. H., & Herman, J. l. (2005). A historical perspective on validity arguments for accountability testing. *Yearbook of the National Society for the Study of Education 104*(2), 1–34. https://doi.org/10.1111/j.1744-7984.2005.00023.x

Hambleton, R. K., & Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. *European Journal of Psychological Assessment*, *11*, 147–157.

Hambleton, R. K., & Lee, M. K. (2013). Methods for translating and adapting tests to increase cross-language validity. In D. H. Saklofske, C. R. Reynolds, & V. Schwean (Eds.), *The Oxford handbook of child psychological assessment* (pp. 172-181). https://doi.org/10.1093/oxfordhb/9780199796304.013.0008

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.). (2004). *Adapting educational and psychological tests for cross-cultural assessment*. Taylor & Francis Group.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). SAGE.

Hanushek, E. A., Peterson, P. E., Talpey, L. M., & Woessmann, L. (2019). The achievement gap fails to close: half century of testing shows persistent divide between haves and have-nots. *Education Next*, *19*(3), 8–18.

Hanushek, E. A., & Rivkin, S. G. (2006). Teacher quality. *Handbook of the Economics of Education*, *2*, 1051-1078.

Harris, D. (1989). Comparison of 1-, 2-, and 3-parameter IRT models. *Educational Measurement: Issues and Practice*, *8*(1), 35–41. https://doi.org/10.1111/j.1745-3992.1989.tb00313.x

Hart, G., & Zinn, S. (2015). The drive for school libraries in South Africa: Intersections and connections. *Library Trends*, *64*(1), 19–41. https://doi.org/10.1353/lib.2015.0035

Haynes, W. (2013). Benjamini–Hochberg method. In W. Dubitzky, O. Wolkenhauer, K. Cho, & H. Yokota (Eds.), *Encyclopedia of Systems Biology* (pp. 78–78). Springer. https://doi.org/10.1007/978-1-4419-9863-7_1215

He, J., Barrera-Pedemonte, F., & Buchholz, J. (2018). Cross-cultural comparability of noncognitive constructs in TIMSS and PISA. *Assessment in Education: Principles, Policy & Practice*, *26*(4), 369–385. https://doi.org/10.1080/0969594X.2018.1469467

Heath, S. B. (1982). What no bedtime story means: Narrative skills at home and school. *Language in Society*, *11*(1), 49–76. https://doi.org/10.1017/S0047404500009039

Heinz, A., Sischka, P. E., Catunda, C., Cosma, A., García-Moya, I., Lyyra, N., Kaman, A., Ravens-Sieberer, U., & Pickett, W. (2022). Item response theory and differential test functioning analysis of the HBSC-Symptom-Checklist across 46 countries. *BMC Medical Research Methodology*, *22*(1). https://doi.org/10.1186/S12874-022-01698-3

Heppt, B., Olczyk, M., & Volodina, A. (2022). Number of books at home as an indicator of socioeconomic status: Examining its extensions and their incremental validity for academic achievement. *Social Psychology of Education*, *25*(4), 903–928. https://doi.org/10.1007/s11218-022-09704-8

Hermann, Z., & Kopasz, M. (2021). Educational policies and the gender gap in test scores: A cross-country analysis. *Research Papers in Education*, *36*(4), 461–482. https://doi.org/10.1080/02671522.2019.1678065

Heugh, K. (2013). Multilingual education policy in South Africa constrained by theoretical and historical disconnections. *Annual Review of Applied Linguistics*, *33*, 215–237. https://doi.org/10.1017/S0267190513000135

Heyneman, S. P., & Loxley, W. A. (1982). Influences on academic achievement across high and low income countries: A re-analysis of IEA data. *Sociology of Education*, *55*(1), 13-21. https://doi.org/10.2307/2112607

Heyneman, S. P., & Loxley, W. A. (1983). The effect of primary-school quality on academic achievement across twenty-nine high- and low-income countries. *American Journal of Sociology*, *88*(6), 1162–1194.

Hoff, E. (2013). Interpreting the early language trajectories of children from low-SES and language minority homes: implications for closing achievement gaps. *Developmental Psychology*, *49*(1), 4–14. https://doi.org/10.1037/a0027238

Hooper, M., Mullis, I. V. S., & Martin, M. O. (2015). PIRLS 2016 context questionnaire framework. In I. V. S. Mullis & M. O. Martin (Eds.), *PIRLS 2016 Assessment Framework*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Hoover, W., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing*, *2*(2), 127–160. https://doi.org/10.1007/BF00401799

Hopfenbeck, T. N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J., & Baird, J.-A. (2018). Lessons learned from PISA: A systematic review of peer-reviewed articles on the Programme for International Student Assessment. *Scandinavian Journal of Educational Research*, *62*(3), 333–353. https://doi.org/10.1080/00313831.2016.1258726

Howard, E. R., Páez, M. M., August, D. L., Barr, C. D., Kenyon, D., & Malabonga, V. (2014). The importance of SES, home and school language and literacy practices, and oral vocabulary in bilingual children's English reading development. *Bilingual Research Journal*, *37*(2), 120–141. https://doi.org/10.1080/15235882.2014.934485

Howie, S. (2010, July 1-3). The relationship between early childhood backgrounds and reading achievement in low and high achieving countries in PIRLS 2006 [Conference presentation]. The 4th IEA International Research Conference, Gothenburg, Sweden.

Howie, S. (2012). High-stakes testing in South Africa: Friend or foe?. *Assessment in Education: Principles, Policy & Practice*, *19*(1), 81–98. https://doi.org/10.1080/0969594X.2011.613369

Howie, S. (2022). Regional studies in non-western countries, and the case of SACMEQ. In T. Nilsen & A. Stancel-Piątak (Eds.), *International handbook of comparative large-scale studies in education: Perspectives, methods and findings* (pp. 420–464). Springer.

Howie, S., Combrinck, C., Roux, K., Mokoena, G. M., & McLeod, P. N. (2017). *South Africa Grade 4 PIRLS Literacy 2016 highlights report: South Africa*. Centre for Evaluation and Assessment, University of Pretoria.

Howie, S., Combrinck, C., Roux, K., Tshele, M., Mokoena, G. M., & McLeod, P. N. (2017). PIRLS Literacy 2016 Progress in International Reading Literacy Study 2016: South African children's reading literacy achievement. Centre for Evaluation and Assessment, University of Pretoria.

Howie, S., Combrinck, C., Tshele, M., Roux, K., Palane, N. M., & Mokoena, G. M. (2017). PIRLS 2016 Progress in International Reading Literacy Study 2016 Grade 5 benchmark participation: South African children's reading literacy achievement. Centre for Evaluation and Assessment, University of Pretoria.

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

Howie, S., Venter, E., & van Staden, S. (2008). The effect of multilingual policies on performance and progression in reading literacy in South African primary schools. *Educational Research and Evaluation, 14*(6), 551-560. https://doi.org/10.1080/13803610802576775

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

Huang, X., Wilson, M., & Wang, L. (2016). Exploring plausible causes of differential item functioning in the PISA science assessment: language, curriculum or culture. *Educational Psychology*, *36*(2), 378–390.

Huey, E. B. (1908). *The psychology and pedagogy of reading*. The Macmillan Company.

Hughes, D. J. (2018). Psychometric validity: Establishing the accuracy and appropriateness of psychometric measures. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (Vol. 2, pp. 759–779). Wiley Blackwell.

Hulme, C., & Snowling, M. J. (2013). Learning to read: What we know and what we need to understand better. *Child Development Perspectives*, *7*(1), 1–5. https://doi.org/10.1111/cdep.12005

Hulme, C., Snowling, M. J., & Nation, K. (Eds.). (2022). *The science of reading: A handbook*. Wiley-Blackwell.

Hunter, C. V. (2014). A simulation study comparing two methods of evaluating differential test functioning (DTF): DFIT and the Mantel-Haenszel/Liu-Agresti Variance [Doctoral dissertation, Georgia State University].

IEA. (2020, March). *IEA Studies: PIRLS*. PIRLS 2016 Results. https://www.iea.nl/studies/iea/pirls/2016/results (Accessed 20 October 2023)

IEA. (2021). Scaling the prePIRLS 2011 South Africa data: August 2021 [unpublished paper].

IEA. (2023). *IEA Studies: LaNA*. Literacy and Numeracy Assessment. https://www.iea.nl/studies/iea/LaNA (Accessed 18 November 2023)

ITC. (2001). International guidelines for test use. *International Journal of Testing*, *1*(2), 93–114.

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

ITC. (2019). ITC guidelines for the large-scale assessment of linguistically and culturally diverse populations. *International Journal of Testing*, *19*(4), 301–336. https://doi.org/10.1080/15305058.2019.1631024

Jacob, B., & Rothstein, J. (2016). The measurement of student ability in modern assessment systems. *Journal of Economic Perspectives*, *30*(3), 85–108. https://doi.org/10.1257/jep.30.3.85

James, Z. (2021). Stakeholder perspectives on medium of instruction policy in Ethiopia. In C. Benson & K. Kosonen (Eds.), *Language Issues in Comparative Education II* (pp. 146–162). Brill. https://doi.org/10.1163/9789004449671_007

Janks, H. (2011). Making sense of the PIRLS 2006 results for South Africa. *Reading & Writing – Journal of the Reading Association of South Africa*, *2*(1), 27–39. https://doi.org/10.4102/rw.v2i1.11

Jansen, J. D. (2019). Inequality in education: What is to be done? In N. Spaull & J. Jansen (Eds.), *South African schooling: The enigma of inequality* (pp. 355–371). Springer.

Jerrim, J. (2011). England's "plummeting" PISA test scores between 2000 and 2009: Is the performance of our secondary school pupils really in relative decline (DoQSS Working Papers No. 09). Quantitative Social Science – UCL Social Research Institute, University College London. https://ideas.repec.org/p/qss/dqsswp/1109.html

Johansson, S. (2016). International large-scale assessments: what uses, what consequences? *Educational Research*, *58*(2), 139–148. https://doi.org/10.1080/00131881.2016.1165559

Johnson, L. L., & Perry, K. H. (2012). What is literacy? A critical overview of sociocultural perspectives. *Journal of Language and Literacy Education*, *8*(1), 50–71.

Joint Education Trust. (2023, May 17). *JET responds to the 2021 PIRLS results* [Press statement]. JET Education Services. https://www.jet.org.za/news/jet-pirls-2021

Jones, L. V, & Thissen, D. (2006). A history and overview of psychometrics. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 1–27). Elsevier. https://doi.org/10.1016/S0169-7161(06)26001-2

Jude, N., Teltemann, J., Parcerisa, L., Fontdevila, C., & Verger, A. (2021). Understanding the PISA influence on national education policies: A focus on policy transfer mechanisms. In A. Wilmers & S. Jornitz (Eds.), *International Perspectives on School Settings, Education Policy and Digital Strategies* (pp. 182–194). Verlag Barbara Budrich. https://doi.org/10.2307/J.CTV1GBRZF4.14

Justman, M. (2018). Randomized controlled trials informing public policy: Lessons from project STAR and class size reduction. *European Journal of Political Economy*, *54*, 167–174. https://doi.org/10.1016/j.ejpoleco.2018.04.005

Kalinowski, S. T. (2019). A graphical method for displaying the model fit of item response theory trace lines. *Educational and Psychological Measurement*, *79*(6), 1064–1074.

Kamens, D. H., & Benavot, A. (2011). National, regional and international learning assessments: Trends among developing countries, 1960–2009. *Globalisation, Societies and Education*, *9*(2), 285–300. https://doi.org/10.1080/14767724.2011.577337

Kamens, D. H., & McNeely, C. L. (2010). Globalization and the growth of international educational testing and national assessment. *Comparative Education Review*, *54*(1), 5–25. https://doi.org/10.1086/648471

Kan, R. T. Y., & Murphy, V. A. (2020). Effects of frequency and idiomaticity on second language reading comprehension in children with English as an additional language. *Studies in Second Language Learning and Teaching*, *10*(3), 579–605.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527–535. https://doi.org/10.1037/0033-2909.112.3.527

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319–342. https://doi.org/10.1111/J.1745-3984.2001.TB01130.X

Kane, M. T. (2013). The argument-based approach to validation. *School Psychology Review*, *42*(4), 448–457. https://doi.org/10.1080/02796015.2013.12087465

Kang, E. Y., & Shin, M. (2019). The contributions of reading fluency and decoding to reading comprehension for struggling readers in fourth grade. *Reading & Writing Quarterly*, *35*(3), 179–192. https://doi.org/10.1080/10573569.2018.1521758

Kanjee, A. (2018). Use of international assessments to support learning at the bottom of the pyramid. In D. A. Wagner, S. Wolf, & R. F. Boruch (Eds.), *Learning at the bottom of the pyramid: Science, measurement, and policy in low-income countries* (pp. 77–92). International Institute for Educational Planning UNESCO.

Kanjee, A., & Moloi, Q. (2014). South African teachers' use of national assessment data. *South African Journal of Childhood Education*, *4*(2), 90–113.

Kanjee, A., & Sayed, Y. (2013). Assessment policy in post-apartheid South Africa: Challenges for improving education quality and learning. *Assessment in Education: Principles, Policy & Practice*, *20*(4), 442–469. https://doi.org/10.1080/0969594X.2013.838541

Kieffer, M. J., Lesaux, N. K., Rivera, M., & Francis, D. J. (2009). Accommodations for English language learners taking large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research*, *79*(3), 1168–1201. https://doi.org/10.3102/0034654309332490

Kim, Y.-S. G. (2017). Why the simple view of reading is not simplistic: Unpacking component skills of reading using a direct and indirect effect model of reading (DIER). *Scientific Studies of Reading*, *21*(4), 310–333. https://doi.org/10.1080/10888438.2017.1291643

Kintsch, W. (2013). Revisiting the construction-integration model of text comprehension and its implications for instruction. In D. Alvermann, N. J. Unrau, & R. B. Ruddell (Eds.), *Theoretical models and processes of literacy* (pp. 807–841). International Reading Association.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363-394. https://doi.org/10.1037/0033-295X.85.5.363

Kirsch, I., Lennon, M., von Davier, M., Gonzalez, E., & Yamamoto, K. (2013). On the growing importance of international large-scale assessments. In M. von Davier, E. Gonzalez, I. Kirsch, & K. Yamamoto (Eds.), *The Role of International Large-Scale Assessments: Perspectives from technology, economy, and educational research* (pp. 1–12). Springer.

Kleinman, M., & Teresi, J. A. (2016). Differential item functioning magnitude and impact measures from item response theory models. *Psychological Test and Assessment Modeling*, *58*(1), 79-91.

Klemenčič, E., & Mirazchiyski, P. V. (2018). League tables in educational evidence-based policy-making: Can we stop the horse race, please? *Comparative Education*, *54*(3), 309–324. https://doi.org/10.1080/03050068.2017.1383082

Klieme, E. (2013). The role of large-scale assessments in research on educational effectiveness and school development. In M. von Davier, E. Gonzalez, I. Kirsch, & K. Yamamoto (Eds.), *The Role of International Large-Scale Assessments: Perspectives from technology, economy, and educational research* (pp. 115–148). Springer.

Klieme, E. (2020). Policies and practices of assessment: A showcase for the use (and misuse) of international large-scale assessments in educational effectiveness research. In J. Hall, A. Lindorff, & P. Sammons (Eds.), *International Perspectives in Educational Effectiveness Research* (pp. 147–181). Springer International Publishing.

Knoch, U., & Chapelle, C. A. (2017). Validation of rating processes within an argument-based framework. *Language Testing*, *35*(4), 477–499. https://doi.org/10.1177/0265532217710049

Koh, A. (2018). "Inter-referencing" as methodology: The "emotional contagion" of PISA and the discursive formation of an emergent East Asia policy field. In K. J. Kennedy & J. C. K. Lee (Eds.), *Routledge international handbook of schools and schooling in Asia* (pp. 40-57). Routledge.

Kotzé, J., Fleisch, B., & Taylor, S. (2019). Alternative forms of early grade instructional coaching: Emerging evidence from field experiments in South Africa. *International Journal of Educational Development*, *66*, 203–213. https://doi.org/10.1016/j.ijedudev.2018.09.004

Kreiner, S., & Christensen, K. B. (2014). Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika*, *79*(2), 210–231.

Kretzer, M. M., & Kaschula, R. H. (2021). Language policy and linguistic landscapes at schools in South Africa. *International Journal of Multilingualism*, *18*(1), 105–127. https://doi.org/10.1080/14790718.2019.1666849

Krijnen, E., van Steensel, R., Meeuwisse, M., & Severiens, S. (2021). Measuring parental literacy beliefs in a socio-economically, linguistically and ethnically diverse sample. *Early Education and Development*, *32*(4), 608-635. https://doi.org/10.1080/10409289.2020.1785227

Kunnan, A. J. (2008). Towards a model of test evaluation: Using the test fairness and the test context frameworks. In L. Taylor & C. J. Weir (Eds.), *Multilingualism and Assessment: Achieving transparency, assuring quality, sustaining diversity* (pp. 229-251). Cambridge University Press.

Kunnan, A. J. (2015). Assessing the quality of large-scale assessments: The case for a fairness and justice approach. In F. A. Hamied, I. B. P. Yadnya, & I. G. A. G. Sosiowati (Eds.), *Developing Indigenous Models of English Language Teaching and Assessment* (pp. 131–141). Udayana University Press.

Kuo, L., & Anderson, R. C. (2006). Morphological awareness and learning to read: A cross-language perspective. *Educational Psychologist*, *41*(3), 161–180. https://doi.org/10.1207/s15326985ep4103_3

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, *6*(2), 293–323. https://doi.org/10.1016/0010-0285(74)90015-2

Lagravinese, R., Liberati, P., & Resce, G. (2020). The impact of economic, social and cultural conditions on educational attainments. *Journal of Policy Modeling*, *42*(1), 112–132. https://doi.org/10.1016/j.jpolmod.2019.03.007

Laitin, D. D., Ramachandran, R., & Walter, S. L. (2019). The legacy of colonial language policies and their impact on student learning: Evidence from an experimental program in Cameroon. *Economic Development and Cultural Change*, *68*(1), 239–272. https://doi.org/10.1086/700617

Landi, N. (2012). Learning to read words: Understanding the relationship between reading ability, lexical quality, and reading context. In M. A. Britt, S. Goldman, & J.-F. Rouet (Eds.), *Reading – from words to multiple texts* (pp. 17–33). Routledge.

Language in Education Policy, Act No. 27 of 1996.

Lantolf, J. P., & Poehner, M. E. (2016). The unfairness of equal treatment: objectivity in L2 testing and dynamic assessment. In H. Karami (Ed.), *Fairness Issues in educational assessment* (pp. 53-69). Routledge.

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension. In C. Laurén & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (pp. 316-340). Multilingual Matters.

Lawrence, M. (2020). Testing 3, 2, 1: What Australian education can learn from Finland. Melbourne Books.

Leibbrandt, M., & Pabón, F. A. D. (2021). Inequality in South Africa. In A. Oqubay, F. Tregenna, & I. Valodia (Eds.), *The Oxford handbook of the South African economy* (pp. 175–194). Oxford University Press.

Lenkeit, J., & Caro, D. H. (2014). Performance status and change – measuring education system effectiveness with data from PISA 2000–2009. *Educational Research and Evaluation*, *20*(2), 146–174. https://doi.org/10.1080/13803611.2014.891462

Lenkeit, J., Chan, J., Hopfenbeck, T. N., & Baird, J.-A. (2015). A review of the representation of PIRLS related research in scientific journals. *Educational Research Review*, *16*, 102–115. https://doi.org/10.1016/j.edurev.2015.10.002

Lenkeit, J., & Schwippert, K. (2018). Doing research with international assessment studies: Methodological and conceptual challenges and ways forward. In *Assessment in Education: Principles, Policy & Practice, 25*(1), 1-4. https://doi.org/10.1080/0969594X.2017.1352137

Lervåg, A., Hulme, C., & Melby-Lervåg, M. (2018). Unpicking the developmental relationship between oral language skills and reading comprehension: It's simple, but complex. *Child Development*, *89*(5), 1821–1838. https://doi.org/10.1111/cdev.12861

Letseka, M. (2014). The illusion of education in South Africa. *Procedia – Social and Behavioral Sciences*, *116*, 4864–4869. https://doi.org/10.1016/j.sbspro.2014.01.1039

Lewis, S., & Lingard, B. (2015). The multiple effects of international large-scale assessment on education policy and research. *Discourse*, *36*(5), 621–637. https://doi.org/10.1080/01596306.2015.1039765

Li, H., Hunter, C. V., & Bialo, J. A. (2022). A revisit of Zumbo's third generation DIF: How are we doing in language testing? *Language Assessment Quarterly*, *19*(1), 27–53. https://doi.org/10.1080/15434303.2021.1963253

Li, Z., Chen, M. Y., & Banerjee, J. (2020). Using corpus analyses to help address the DIF interpretation: Gender differences in standardized writing assessment. *Frontiers in psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.01088

Lin, M., Bumgarner, E., & Chatterji, M. (2014). Understanding validity issues in international large scale assessments. *Quality Assurance in Education*, *22*(1), 31–41. https://doi.org/10.1108/qae-12-2013-0050

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*(1), 85–106.

Lindorff, A., Stiff, J. C., & Kayton, H. L. (2023). *PIRLS 2021: National report for England*. Department for Education & University of Oxford.

Lindquist, E. F. (Ed.). (1951). *Educational measurement*. American Council on Education.

Little, T. D., Donnellan, M. B., & Lucas, R. E. (2013). Secondary data analysis. In T. D. Little (Ed.), *The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2 Statistical Analysis* (pp. 665-677). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199934898.013.0028

Lockheed, M. E., & Wagemaker, H. (2013). International large-scale assessments: Thermometers, whips or useful policy tools? *Research in Comparative and International Education*, *8*(3). https://doi.org/10.2304/rcie.2013.8.3.296

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Routledge.

Lumley, T., & Scott, A. (2015). AIC and BIC for modeling with complex survey data. *Journal of Survey Statistics and Methodology*, *3*(1), 1–18. https://doi.org/10.1093/jssam/smu021

Luo, R., Tamis-LeMonda, C. S., & Mendelsohn, A. L. (2020). Children's literacy experiences in low-income families: The content of books matters. *Reading Research Quarterly*, *55*(2), 213–233. https://doi.org/10.1002/RRQ.263

Maimane, M. (2023, May 22). PIRLS results are a blaring fire alarm for SA society. *The Daily Maverick*. https://www.dailymaverick.co.za/opinionista/2023-05-22-pirls-results-a-blaring-fire-alarm-for-south-african-society/

Mair, P. (2018). *Modern psychometrics with R*. Springer International Publishing. https://doi.org/10.1007/978-3-319-93177-7

Makalela, L. (2015). Translanguaging principles in L2 reading instruction: Implications for ESL pre-service teacher programme. In T. S. C. Farrell (Ed.), *International Perspectives on English Language Teacher Education* (pp. 107–124). https://doi.org/10.1057/9781137440068_7

Makalela, L. (2017). Bilingualism in South Africa: Reconnecting with ubuntu translanguaging. In O. García, A. M. Y. Lin, & S. May (Eds.), *Bilingual and Multilingual Education* (pp. 297–301). Springer.

Makalela, L. (2016). Ubuntu translanguaging: An alternative framework for complex multilingual encounters. *Southern African Linguistics and Applied Language Studies*, 34(3), 187-196.

Makalela, L. (2023). Merging English Home Language and First Additional Language curricula: Implications for future quality assurance practices. *Southern African Linguistics and Applied Language Studies*, *41*(1), 76–87. https://doi.org/10.2989/16073614.2023.2185984

Makaluza, N. (2020). *isiXhosa text analysis*. Funda Wande. https://fundawande.org/news/isixhosa-text-analysis--nwabisa-makaluza-9 (Accessed 22 January 2021).

Malda, M., Nel, C., & van de Vijver, F. J. R. (2014). The road to reading for South African learners: The role of orthographic depth. *Learning and Individual Differences*, *30*, 34–45.

Marks, G. N. (2008). Accounting for the gender gaps in student performance in reading and mathematics: Evidence from 31 countries. *Oxford Review of Education*, *34*(1), 89–109. https://doi.org/10.1080/03054980701565279

Marôco, J. (2021). What makes a good reader? Worldwide insights from PIRLS 2016. *Reading and Writing*, *34*(1), 231–272. https://doi.org/10.1007/S11145-020-10068-8

Martin, M. O., Mullis, I. V. S., & Foy, P. (2017). Assessment design for PIRLS, PIRLS Literacy, and ePIRLS in 2016. In I. V. S. Mullis & M. O. Martin (Eds.), *PIRLS 2016 Assessment Framework*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Martin, M. O., Mullis, I. V. S., & Hooper, M. (2017). *Methods and procedures in PIRLS 2016*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Martin, M. O., von Davier, M., Foy, P., & Mullis, I. V. S. (2019). PIRLS 2021 assessment design. In I. V. S. Mullis & M. O. Martin (Eds.), *PIRLS 2021 Assessment frameworks*. Boston College, TIMSS & PIRLS International Study Center.

Maul, A., Mari, L., & Wilson, M. (2019). Intersubjectivity of measurement across the sciences. *Measurement*, *131*, 764–770. https://doi.org/10.1016/j.measurement.2018.08.068

Maynier, D. (2023, May 16). *PIRLS 2021 results: We have a plan to get reading scores #BackOnTrack* [MEC Statement on Progress in International Reading and Literacy Study 2021 Results]. South African Government.

McBride, C. (2017). Early literacy across languages. In R. K. Parrila, K. Cain, & D. L. Compton (Eds.), *Theories of reading development* (pp. 55–74). John Benjamins Publishing Company.

McClung, N. A., & Pearson, P. D. (2019). Reading comprehension across languages: Seven European orthographies and two international literacy assessments. *Written Language & Literacy*, *22*(1), 33–66.

McGrane, J. A., & Maul, A. (2020). The human sciences, models and metrological mythology. *Measurement*, *152*, 107346. https://doi.org/10.1016/j.measurement.2019.107346

McKeown, M. G., Beck, I. L., & Blake, R. G. K. (2009). Rethinking reading comprehension instruction: A comparison of instruction for strategies and content approaches. *Reading Research Quarterly*, *44*(3), 218–253.

McKinney, C. (2017). Language and power in post-colonial schooling: Ideologies in practice. Routledge.

McKinney, C. (2024). Coloniality of language and pretextual gaps: a case study of emergent bilingual children’s writing in a South African school and a call for ukuzilanda. *Journal of Multilingual and Multicultural Development*, *45*(3), 663-679.

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology*, *95*(4), 728–743. https://doi.org/10.1037/A0018966

Meade, A. W., & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *Journal of Applied Psychology*, *97*(5), 1016–1031. https://doi.org/10.1037/A0027934

Melby-Lervåg, M., & Lervåg, A. (2014). Reading comprehension and its underlying components in second-language learners: A meta-analysis of studies comparing first- and second-language learners. *Psychological Bulletin*, *140*(2), 409–433. https://doi.org/10.1037/A0033890

Melby-Lervåg, M., Lyster, S. A. H., & Hulme, C. (2012). Phonological skills and their role in learning to read: A meta-analytic review. *Psychological Bulletin*, *138*(2), 322–352. https://doi.org/10.1037/A0026744

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*(11), 1012–1027. https://doi.org/10.1037/0003-066X.35.11.1012

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11. https://doi.org/10.3102/0013189X018002005

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research*, *45*(1–3), 35–44. https://doi.org/10.1023/A:1006964925094

Meyer, H.-D., & Zahedi, K. (2014, May 6). Open letter: OECD and Pisa tests are damaging education worldwide. *The Guardian*. https://www.theguardian.com/education/2014/may/06/oecd-pisa-tests-damaging-education-academics

Michaelides, M. P., & Haertel, E. H. (2004). *Sampling of common items: An unrecognized source of error in test equating* (CSE Report No. 636). National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Miller, D., Bayram, F., Rothman, J., & Serratrice, L. (2018). *Bilingual cognition and language: The state of the science across its subfields* (Vol. 54). John Benjamins Publishing Company.

Miller, S. D., & Faircloth, B. S. (2009). Motivation and reading comprehension. In S. E. Israel & G. G. Duffy (Eds.), *Handbook of research on reading comprehension* (pp. 227–239). Routledge.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*(2), 177–196. https://doi.org/10.1007/BF02294457

Mohangi, K., Krog, S., Stephens, O., & Nel, N. (2016). Contextual challenges in early literacy teaching and learning in Grade R rural schools in South Africa. *Per Linguam*, *32*(1). https://doi.org/10.5785/32-1-646

Mohohlwane, N. L. (2019). How language policy and practice sustains inequality in education. In N. Spaull & J. D. Jansen (Eds.), *South African schooling: The enigma of inequality* (pp. 127–146). Springer.

Moos, L. (2014). Educational governance in Denmark. *Leadership and Policy in Schools*, *13*(4), 424–443. https://doi.org/10.1080/15700763.2014.945655

Motala, S., & Carel, D. (2019). Educational funding and equity in South African schools. In N. Spaull & J. Jansen (Eds.), *South African schooling: The enigma of inequality* (pp. 25–45). Springer.

Motshekga, A. (2021, August 24). Opening address by the minister of basic education Angie Motshekga. *South African Human Rights Commission's Right to Read and Write Campaign Launch*.

Mphasha, L. E. (2016). Language identity and the development of multilingualism in South Africa. *Journal of Sociology and Social Anthropology*, *7*(1), 27–34. https://doi.org/10.1080/09766634.2016.11885699

Mthiyane, N. (2016). Pre-Service Teachers' Beliefs and Experiences Surrounding the Use of Language in Science Classrooms: A South African Case Study. *Nordic Journal of African Studies*, 25(2), 19. https://doi.org/10.53228/njas.v25i2.112

Mtsatse, N., & van Staden, S. (2021). Exploring differential item functioning on reading achievement: A case for isiXhosa. *South African Journal of African Languages*, *41*(1), 1–11. https://doi.org/10.1080/02572117.2021.1902127

Mullis, I. V. S., Martin, M. O., Foy, P., & Drucker, K. (2012). *PIRLS 2011 international results in reading*. TIMSS & PIRLS International Study Center, Boston College. https://timssandpirls.bc.edu/pirls2011/international-results-pirls.html

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017a). PIRLS 2016 international results in reading. In *International Association for the Evaluation of Educational Achievement*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017b). What makes a good reader: International findings from PIRLS 2016. In Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M., *PIRLS 2016 International Results in Reading*. https://pirls2016.org/pirls/summary/

Mullis, I. V. S., Martin, M. O., & Kennedy, A. M. (2009). *PIRLS 2011 assessment framework*. TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., & Foy, P. (2007). *PIRLS 2006 International report: IEA's Progress in International Reading Literacy Study in Primary Schools in 40 Countries*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Mullis, I. V. S., Martin, M. O., & Sainsbury, M. (2015a). *PIRLS 2016 Assessment Framework*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Mullis, I. V. S., Martin, M. O., & Sainsbury, M. (2015b). PIRLS 2016 reading framework. In I. V. S. Mullis & M. O. Martin (Eds.), *PIRLS 2016 Assessment Framework*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Mullis, I. V. S., von Davier, M., Foy, P., Fishbein, B., Reynolds, K. A., & Wry, E. (2023). *PIRLS 2021 international results in reading*. Boston College, TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.tr2103.kb5342

Mulongo, G., & Amod, Z. (2019). Nexus between participating in cross-national learning assessments and educational policy reforms: Case study of Kenya, Tanzania and South Africa. *Educational Research for Policy and Practice*, *18*(3), 203–223. https://doi.org/10.1007/s10671-019-09250-9

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176. https://doi.org/10.1177/014662169201600206

Muraki, E. (1997). A generalized partial credit model. In W. J. Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 153–164). https://doi.org/10.1007/978-1-4757-2691-6_9

Murphy, V. A. (2018). Literacy development in linguistically diverse pupils. In D. Miller, F. Bayram, J. Rothman, & L. Serratrice (Eds.), *Bilingual cognition and language: The state of the science across its subfields. Studies in Bilingualism* (Vol. 54, pp. 155–182). John Benjamins.

Murphy, V. A., & Unthiah, A. (2015). A systematic review of intervention research examining English language and literacy development in children with English as an Additional Language (EAL). Education Endowment Foundation, Unbound, The Bell Foundation & University of Oxford.

Mutekwe, E., & Sedibe, M. (2015). South Africa's quest for democratic education through a language in education policy perspective. *Creative Education*, *6*, 1076–1085.

Muter, V., Hulme, C., Snowling, M. J., & Stevenson, J. (2004). Phonemes, rimes, vocabulary, and grammatical skills as foundations of early reading development: Evidence from a longitudinal study. *Developmental Psychology*, *40*(5), 665–681. https://doi.org/10.1037/0012-1649.40.5.665

Nag, S., Vagh, S. B., Dulay, K. M., Snowling, M. J., Donolato, E., & Melby-Lervåg, M. (2023). Home learning environments and children's language and literacy skills: A meta-analytic review of studies conducted in low- and middle-income countries. *Psychological Bulletin*. https://doi.org/10.17605/osf.io/r9bgv

Naidoo, K. (1990). The politics of student resistance in the 1980s. In M. Nkomo (Ed.), *Pedagogy of domination: Toward a democratic education in South Africa* (pp. 121–145). Africa World Press.

Naidoo, U., Reddy, K., & Dorasamy, N. (2014). Reading literacy in primary schools in South Africa: Educator perspectives on factors affecting reading literacy and strategies for improvement. *International Journal of Educational Sciences*, *7*(1), 155–167.

Nakidien T, Singh M, & Sayed Y. (2021). Teachers and Teacher Education: Limitations and Possibilities of Attaining SDG 4 in South Africa. *Education Sciences*, 11(2), 66. https://doi.org/10.3390/educsci11020066

National Education Policy, Act No. 27 of 1996.

National Planning Commission. (2010). *National Development Plan – 2030*. South African Government.

Ndimande, B. S. (2013). From Bantu education to the fight for socially just education. *Equity and Excellence in Education 46*(1), 20-35. https://doi.org/10.1080/10665684.2013.750199

Nel, N., Krog, S., Mohangi, K., Muller, H., & Stephens, O. (2016). Research partnership between South Africa and China: Emergent literacy teaching and learning in early childhood education in South Africa. *Per Linguam*, *32*(1). https://doi.org/10.5785/32-1-647

Nel, N., & Müller, H. (2010). The impact of teachers' limited English proficiency on English second language learners in South African schools. *South African Journal of Education*, *30*(4), 635-650.

Netten, A., Luyten, H., Droop, M., & Verhoeven, L. (2016). Role of linguistic and sociocultural diversity in reading literacy achievement: A multilevel approach. *Journal of Research in Reading*, *39*(2), 189–208. https://doi.org/10.1111/1467-9817.12032

Netten, A., Voeten, M., Droop, M., & Verhoeven, L. (2014). Sociocultural and educational factors for reading literacy decline in the Netherlands in the past decade. *Learning and Individual Differences*, *32*, 9–18. https://doi.org/10.1016/j.lindif.2014.02.002

Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy & Practice*, *14*(2), 149–170. https://doi.org/10.1080/09695940701478321

Newton, P. E. (2012). Clarifying the consensus definition of validity. *Measurement: Interdisciplinary Research and Perspectives*, *10*(1–2), 1–29. https://doi.org/10.1080/15366367.2012.669666

Newton, P. E. (2017). There is more to educational measurement than measuring: The importance of embracing purpose pluralism. *Educational Measurement: Issues and Practice*, *36*(2), 5–15. https://doi.org/10.1111/emip.12146

Newton, P. E., & Shaw, S. D. (2014). Validity in educational and psychological assessment. SAGE.

Ngobeni, N. R., Chibambo, M. I., & Divala, J. J. (2023). Curriculum transformations in South Africa: Some discomforting truths on interminable poverty and inequalities in schools and society. *Frontiers in Education*, *8*, 1132167.

Ngubane, N. I., & Makua, M. (2022). Intersection of Ubuntu pedagogy and social justice: Transforming South African higher education. *Transformation in Higher Education, 6*, a113. https://doi.org/10.4102/the.v6i0.113

Nilsen, T., Stancel-Piątak, A., & Gustafsson, J.-E. (Eds.). (2022). International handbook of comparative large-scale studies in education: Perspectives, methods and findings. Springer.

Nisbet, I. (2019). Fairness takes centre stage. *Assessment in Education: Principles, Policy & Practice*, *26*(1), 111–117. https://doi.org/10.1080/0969594X.2017.1358151

Nisbet, I., & Shaw, S. D. (2019). Fair assessment viewed through the lenses of measurement theory. *Assessment in Education: Principles, Policy & Practice*, *26*(5), 612–629. https://doi.org/10.1080/0969594X.2019.1586643

Nkadimeng, S., Lepota, B., & Makalela, L. (2023). The language enigma: Two decades of quality assuring language assessment in South Africa. *Southern African Linguistics and Applied Language Studies*, *41*(1), 1–4. https://doi.org/10.2989/16073614.2023.2185966

Nkadimeng, S., & Makalela, L. (2023). The (re)making of an African language: Revisiting epistemologies for quality assessment practices. *Southern African Linguistics and Applied Language Studies*, *41*(1), 5–15. https://doi.org/10.2989/16073614.2023.2185973

Nugraha, S. I. (2018, November 27-28). *The Language-in-Education Policy in South Africa: A gap between policy and efficacy* [Conference presentation]. CONAPLIN 2018: 11th Conference on Applied Linguistics, Bandung, Indonesia.

Nye, C. D., Joo, S. H., Zhang, B., & Stark, S. (2019). Advancing and evaluating IRT model data fit indices in organizational research. *Organizational Research Methods*, *23*(3), 457–486. https://doi.org/10.1177/1094428119833158

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality and Quantity*, *41*(5), 673–690. https://doi.org/10.1007/S11135-006-9018-6

OECD. (2019). PISA 2021 translation and adaptation guidelines. cApStAn.

OECD. (2023). PISA 2022 assessment and analytical framework. OECD Publishing.

Oliveri, M. E., Ercikan, K., & Simon, M. (2015). A framework for developing comparable multilingual assessments for minority populations: Why context matters. *International Journal of Testing*, *15*(2), 94–113. https://doi.org/10.1080/15305058.2014.986271

Oliveri, M. E., Ercikan, K., & Zumbo, B. D. (2013). Analysis of sources of latent class differential item functioning in international assessments. *International Journal of Testing*, *13*(3), 272–293.

Oliveri, M. E., Rutkowski, D., & Rutkowski, L. (2018). Bridging validity and evaluation to match international large-scale assessment claims and country aims. *ETS Research Report Series*, *2018*(1), 1–9. https://doi.org/10.1002/ETS2.12214

Oliveri, M. E., & von Davier, A. (2016). Psychometrics in support of a valid assessment of linguistic minorities: Implications for the test and sampling designs. *International Journal of Testing*, *16*(3), 220–239. https://doi.org/10.1080/15305058.2015.1069743

Omidire, M. F., & Ayob, S. (2022). The utilisation of translanguaging for learning and teaching in multilingual primary classrooms. *Multilingua*, 41(1), 105-129.

Opare-Kumi, J. (2024). English medium instruction in multilingual contexts: Empirical evidence from Ethiopia. *International Journal of Educational Development*, *105*, 102987. https://doi.org/10.1016/j.ijedudev.2024.102987

Ortiz, S. O., Piazza, N., Ochoa, S. H., & Dynda, A. M. (2018). Testing with culturally and linguistically diverse populations: New directions in fairness and validity. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 684–712). The Guilford Press.

Osborne, J. W., & Waters, E. (2019). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research, and Evaluation*, *8*(1), 2. https://doi.org/10.7275/r222-hv23

Oshima, T. C., Raju, N. S., Flowers, C. P., & Slinde, J. A. (1998). Differential bundle functioning using the DFIT framework: Procedures for identifying possible sources of differential functioning. *Applied Measurement in Education*, *11*(4), 353–369. https://doi.org/10.1207/s15324818ame1104_4

Osteen, P. (2010). An introduction to using multidimensional item response theory to assess latent factor structures. *Journal of the Society for Social Work and Research*, *1*(2), 66–82. https://doi.org/10.5243/jsswr.2010.6

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning*. SAGE.

Ouellette, G., & Beers, A. (2010). A not-so-simple view of reading: How oral vocabulary and visual-word recognition complicate the story. *Reading and Writing*, *23*(2), 189–208. https://doi.org/10.1007/s11145-008-9159-1

Ozga, J. (2012). Assessing PISA. *European Educational Research Journal*, *11*(2), 166–171. https://doi.org/10.2304/eerj.2012.11.2.166

Pae, H. K., & Sevcik, R. A. (2011). The role of verbal working memory in second language reading fluency and comprehension: A comparison of English and Korean. *International Electronic Journal of Elementary Education*, *4*(1), 47–65.

Paris, S. G., & Stahl, S. A. (Eds.). (2005). *Children's reading comprehension and assessment*. Routledge. https://doi.org/10.4324/9781410612762

Park, H. (2008). Home literacy environments and children's reading performance: A comparative study of 25 countries. *Educational Research and Evaluation*, *14*(6), 489–505. https://doi.org/10.1080/13803610802576734

Parliament of the Republic of South Africa. (2023, May 3). *The NA approves South African Sign Language as the 12th official language* [Press Release]. Republic of South Africa.

Peaker, G. F. (1975). *An empirical study of education in twenty-one countries: A technical report*. International studies in evaluation.

Pearson, P. D. (2004). The reading wars. *Educational Policy*, *18*(1), 216–252. https://doi.org/10.1177/0895904803260041

Pearson, P. D. (2011). The roots of reading comprehension instruction. In S. E. Israel & G. G. Duffy (Eds.), *Handbook of research on reading comprehension* (Vol. 2, pp. 3–30). Routledge.

Pearson, P. D., & Hamm, D. N. (2005). The assessment of reading comprehension: A review of practices-past, present, and future. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 13–69). Lawrence Erlbaum Associates Publishers.

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review*, *18*(4), 315–341. https://doi.org/10.1007/S10648-006-9029-9

Peña, E. A., & Slate, E. H. (2006). Global validation of linear model assumptions. *Journal of the American Statistical Association*, *101*(473), 341–354. https://doi.org/10.1198/016214505000000637

Penfield, R. D., & Camilli, G. (2006). Differential item functioning and item bias. In C. R. Rao & S. B. T. Sinharay (Eds.), *Handbook of Statistics: Vol. 26 Psychometrics* (pp. 125–167). Elsevier. https://doi.org/10.1016/S0169-7161(06)26005-X

Perez, B., McCarty, T. L., Watahomigie, L. J., & thi Dien, T. (Eds.). (2004). *Sociocultural contexts of language and literacy*. Routledge.

Perfetti, C. A., & Goldman, S. R. (1976). Discourse memory and reading comprehension skill. *Journal of Verbal Learning and Verbal Behavior*, *15*(1), 33–42. https://doi.org/10.1016/S0022-5371(76)90004-9

Perfetti, C. A., & Harris, L. N. (2013). Universal reading processes are modulated by language and writing system. *Language Learning and Development*, *9*(4), 296–316. https://doi.org/10.1080/15475441.2013.813828

Piper, B., Zuilkowski, S. S., Dubeck, M., Jepkemei, E., & King, S. J. (2018). Identifying the essential ingredients to literacy and numeracy improvement: Teacher professional development and coaching, student textbooks, and structured teachers' guides. *World Development*, *106*, 324-336.

Podhajski, B., Mather, N., Nathan, J., & Sammons, J. (2009). Professional development in scientifically based reading instruction. *Journal of Learning Disabilities*, *42*(5), 403–417. https://doi.org/10.1177/0022219409338737

Popat, S., Lenkeit, J., & Hopfenbeck, T. N. (2017). *PIRLS for teachers – A review of practitioner engagement with international large-scale assessment results*. Oxford University Centre for Educational Assessment.

Popham, W. J. (1999). Why standardized tests don't measure educational quality. *Educational Leadership*, *56*, 8-16.

Pretorius, E. J. (2014). Supporting transition or playing catch-up in Grade 4? Implications for standards in education and training. *Perspectives in Education*, *31*(1).

Pretorius, E. J., & Mampuru, D. M. (2007). Playing football without a ball: Language, reading and academic performance in a high-poverty school. *Journal of Research in Reading*, *30*(1), 38–58. https://doi.org/10.1111/j.1467-9817.2006.00333.x

Pretorius, E. J., & Roux, K. (2023). When comprehension barely occurs: Examining low performance in South Africa to better support and monitor reading comprehension. In S. van Staden & C. Combrinck (Eds.), *Tracking changes in South African Reading literacy achievement: A developing context perspective* (pp. 88–120). Brill.

Pretorius, E. J., & Spaull, N. (2016). Exploring relationships between oral reading fluency and reading comprehension amongst English second language readers in South Africa. *Reading and Writing*, *29*(7), 1449–1471. http://dx.doi.org/10.1007/s11145-016-9645-9

Probert, T. N., & De Vos, M. (2016). Word recognition strategies amongst isiXhosa/English bilingual learners: The interaction of orthography and language of learning and teaching. *Reading & Writing – Journal of the Reading Association of South Africa*, *7*(1), 1–10.

Pym, A. (2009). *Exploring Translation Theories*. Routledge. https://doi.org/10.4324/9780203869291

Rajaram, M. C. (2018). *The influence of phonological similarity on multisyllabic word acquisition in children* [Doctoral dissertation, University of Texas]. ProQuest Dissertations Publishing.

Rajlic, G. (2019). *Violations of unidimensionality and local independence in measures intended as unidimensional: Assessing levels of violations and the accuracy in unidimensional IRT model estimates* [Doctoral dissertation, University of British Columbia]. https://doi.org/10.14288/1.0380235

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, *19*(4), 353–368. https://doi.org/10.1177/014662169501900405

Rakgogo, T. J., & Zungu, E. B. (2022). A blatant disregard of Section 6 (1) of the Constitution of South Africa by higher education institutions and language authorities: An onomastic discrepancy. *Literator*, *43*(1), 1–9. https://doi.org/10.4102/lit.v43i1.1814

Ramaphosa, P. C. (2019). *State of the nation address*. Parliament of the Republic of South Africa.

Rappleye, J., Komatsu, H., Uchida, Y., Krys, K., & Markus, H. (2020). 'Better policies for better lives'?: Constructive critique of the OECD's (mis)measure of student well-being. *Journal of Education Policy*, *35*(2), 258–282. https://doi.org/10.1080/02680939.2019.1576923

Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests*. MESA Press.

Reagan, T. G. (1987). The politics of linguistic apartheid: Language policies in black education in South Africa. *The Journal of Negro Education*, *56*(3), 299–312. https://doi.org/10.2307/2295225

Reddy, V. (2010). Cross-national achievement studies: Learning from South Africa's participation in the Trends in International Mathematics and Science Study (TIMSS). *Compare: A Journal of Comparative and International Education*, *35*(1), 63–77. https://doi.org/10.1080/03057920500033571

Reise, S. P. (1990). A comparison of item- and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement*, *14*(2), 127–137. https://doi.org/10.1177/014662169001400202

Reutzel, D. R., Child, A., Jones, C. D., & Clark, S. K. (2014). Explicit instruction in core reading programs. *The Elementary School Journal*, *114*(3), 406–430. https://doi.org/10.1086/674420

Reynolds, C. R., Altmann, R. A., & Allen, D. N. (2021). *Mastering Modern Psychological Testing*. Springer. https://doi.org/10.1007/978-3-030-59455-8_15

Ritzhaupt, A. D., Liu, F., Dawson, K., & Barron, A. E. (2013). Differences in student information and communication technology literacy based on socio-economic status, ethnicity, and gender: Evidence of a digital divide in Florida schools. *Journal of Research on Technology in Education*, *45*(4), 291–307.

Robin, F., Sireci, S. G., & Hambleton, R. K. (2003). Evaluating the equivalence of different language versions of a credentialing exam. *International Journal of Testing*, *3*(1), 1–20. https://doi.org/10.1207/S15327574IJT0301_1

Roever, C., & McNamara, T. (2006). Language testing: The social dimension. *International Journal of Applied Linguistics*, *16*(2), 242–258.

Rojas-Torres, L., Ordóñez, G., & Calvo, K. (2021). Teacher and student practices associated with performance in the PISA reading literacy evaluation. *Frontiers in Education*, *6*, 167.

Roth, W.-M., Oliveri, M. E., Sandilands, D. A., Lyons-Thomas, J., & Ercikan, K. (2013). Investigating linguistic sources of differential item functioning using expert think-aloud protocols in science achievement tests. *International Journal of Science Education*, *35*(4), 546–576. https://doi.org/10.1080/09500693.2012.721572

Roux, K., Mtsatse, N., & Long, C. (2018). A comparison of text difficulty in systemic assessment using Lexile theory. *Southern African Linguistics and Applied Language Studies*, *36*(4), 291–300. https://doi.org/10.2989/16073614.2018.1548290

Roux, K., van Staden, S., & Pretorius, E. J. (2022). Investigating the differential item functioning of a PIRLS Literacy 2016 text across three languages. *Journal of Education*, *87*, 135–155. https://doi.org/10.17159/2520-9868/I87A07

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley. https://doi.org/10.1002/9780470316696

Rueda, R. (2013). 21st-century skills: Cultural, linguistic, and motivational perspectives. In D. Alvermann, N. Unrau, & R. Ruddell (Eds.), *Theoretical models and processes of reading* (pp. 1068–1241). International Reading Association.

Rumelhart, D. E. (1977). Toward an interactive model of reading. In S. Dornic (Ed.), *Attention and performance* (Vol. 6, pp. 573–603). Routledge.

Rutkowski, D., & Rutkowski, L. (2021). Running the wrong race? The case of PISA for Development. *Comparative Education Review*, *65*(1), 147–165. https://doi.org/10.1086/712409

Rutkowski, D., & Rutkowski, L. (2022). Designing measurement for all students in ILSAs. In T. Nilsen, A. Stancel-Piątak, & J.-E. Gustafsson (Eds.), *International Handbook of Comparative Large-Scale Studies in Education* (pp. 685–700). Springer. https://doi.org/10.1007/978-3-030-88178-8_27

Rutkowski, D., Thompson, G., & Rutkowski, L. (2020). Understanding the policy influence of international large-scale assessments in education. In H. Wagemaker (Ed.), *Reliability and validity of international large-scale assessment: Understanding IEA's comparative studies of student achievement* (pp. 261–275). Springer.

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Education Researcher*, *39*(2), 142–151. https://doi.org/10.3102/0013189X10363170

Rutkowski, L., Rutkowski, D., & Liaw, Y.-L. (2019). The existence and impact of floor effects for low-performing PISA participants. *Assessment in Education: Principles, Policy & Practice*, *26*(6), 643–664. https://doi.org/10.1080/0969594X.2019.1577219

Rutkowski, L., Rutkowski, D., & Svetina Valdivia, D. (2022). Multistage test design considerations in international large-scale assessments of educational achievement. In T. Nielsen, A. Stancel-Piątak, & J.-E. Gustafsson (Eds.), *International Handbook of Comparative Large-Scale Studies in Education* (pp. 749–767). Springer. https://doi.org/10.1007/978-3-030-88178-8_63

Rutkowski, L., Rutkowski, D., & Zhou, Y. (2016). Item calibration samples and the stability of achievement estimates and system rankings: Another look at the PISA model. *International Journal of Testing*, *16*(1), 1–20. https://doi.org/10.1080/15305058.2015.1036163

Rutkowski, L., von Davier, M., & Rutkowski, D. (Eds.). (2013). *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*. CRC Press.

SACMEQ. (2024, January 1). *SEACMEQ: The Southern and Eastern Africa Consortium for Monitoring Educational Quality*. SEACMEQ V: Project Description and Implementation. http://www.seacmeq.org/?q=seacmeq-projects/seacmeq-v (Accessed 12 January 2024).

SADTU. (2014). Annual National Assessment (ANA): A SADTU perspective. South African Democratic Teachers Union.

SAHRC. (2021). *The right to read and write*. South African Human Rights Commission.

SAHRC. (2023, May 16). *Progress in International Reading Literacy Study (PIRLS) 2021 results* [Media statement]. South African Human Rights Commission. https://www.sahrc.org.za/index.php/sahrc-media/news-2/item/3589-media-statement-progress-in-international-reading-literacy-study-pirls-2021-results

Salisbury, T. (2016). Education and inequality in South Africa: Returns to schooling in the post-apartheid era. *International Journal of Educational Development*, *46*, 43–52. https://doi.org/10.1016/j.ijedudev.2015.07.004

Sälzer, C., & Prenzel, M. (2014). Looking back at five rounds of PISA: Impacts on teaching and learning in Germany. In M. Štraus (Ed.), *Evidence from the PISA study on educational quality in Slovenia and other countries* (pp. 53–72). Šolsko polje.

Sandefur, J. (2018). Internationally comparable mathematics scores for fourteen African countries. *Economics of Education Review*, *62*, 267–286. https://doi.org/10.1016/j.econedurev.2017.12.003

Sandilands, D. A., Oliveri, M. E., Zumbo, B. D., & Ercikan, K. (2013). Investigating sources of differential item functioning in international large-scale assessments using a confirmatory approach. *International Journal of Testing*, *13*(2), 152–174. https://doi.org/10.1080/15305058.2012.690140

Sarroub, L., & Pearson, P. D. (1998). Two steps forward, three steps back: The stormy history of reading comprehension assessment. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, *72*(2), 97–105. https://doi.org/10.1080/00098659809599604

Schildkamp, K. (2019). Data-based decision-making for school improvement: Research insights and gaps. *Educational Research*, *61*(3), 257–273. https://doi.org/10.1080/00131881.2019.1625716

Schneider, L., Strobl, C., Zeileis, A., & Debelak, R. (2022). An R toolbox for score-based measurement invariance tests in IRT models. *Behavior Research Methods*, *54*(5), 2101–2113. https://doi.org/10.3758/S13428-021-01689-0

Schöfl, M., Steinmair, G., Zepnik, S., & Gnambs, T. (2023). Equivalence of computer- and paper-based administrations of the SCHNAPP spelling test in German for six-year-old children. *European Journal of Psychological Assessment*. https://doi.org/10.1027/1015-5759/A000793

Scholtz, S. E. (2021). Sacrifice is a step beyond convenience: A review of convenience sampling in psychological research in Africa. *SA Journal of Industrial Psychology*, *47*(1), 1–12. https://doi.org/10.4102/sajip.v47i0.1837

Sefotho, M. P., & Makalela, L. (2017). Translanguaging and orthographic harmonisation: A cross-lingual reading literacy in a Johannesburg school. *Southern African Linguistics and Applied Language Studies*, *35*(1), 41–51. https://doi.org/10.2989/16073614.2016.1272423

Seid, Y. (2019). The impact of learning first in mother tongue: evidence from a natural experiment in Ethiopia. *Applied Economics*, *51*(6), 577–593. https://doi.org/10.1080/00036846.2018.1497852

Sen, S. (2018). Spurious latent class problem in the mixed Rasch model: A comparison of three maximum likelihood estimation methods under different ability distributions. *International Journal of Testing*, *18*(1), 71–100. https://doi.org/10.1080/15305058.2017.1312408

Sénéchal, M., & Lefevre, J. A. (2014). Continuity and change in the home literacy environment as predictors of growth in vocabulary and reading. *Child Development*, *85*(4), 1552–1568. https://doi.org/10.1111/cdev.12222

Sénéchal, M., Ouellette, G., & Rodney, D. (2006). The misunderstood giant: On the predictive role of early vocabulary to future reading. In D. K. Dickinson & S. B. Neuman (Eds.), *Handbook of Early Literacy Research* (pp. 173–182). Guilford Press.

Shaw, S. D., & Imam, H. (2013). Assessment of international students through the medium of English: Ensuring validity and fairness in content-based examinations. *Language Assessment Quarterly*, *10*(4), 452–475. https://doi.org/10.1080/15434303.2013.866117

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 159–194. https://doi.org/10.1007/BF02294572

Shen, C., & Tam, H. P. (2008). The paradoxical relationship between student achievement and self-perception: A cross-national analysis based on three waves of TIMSS data. *Educational Research and Evaluation*, *14*(1), 87–100. https://doi.org/10.1080/13803610801896653

Shepherd, D. L. (2018). Understanding language in education and Grade 4 reading performance using a 'natural experiment' of Botswana and South Africa. *Assessment in Education: Principles, Policy & Practice*, *25*(1), 87–106. https://doi.org/10.1080/0969594X.2016.1228604

Sindi-Leigh, M. (2019, April 26). Why are South African children struggling to read properly?. *Africa Is a Country*. https://africasacountry.com/2019/04/why-are-south-african-children-struggling-to-read-properly

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher*, *36*(8), 477–481. https://doi.org/10.3102/0013189X07311609

Sireci, S. G., & Rios, J. A. (2013). Decisions that make a difference in differential item functioning. *Educational Research and Evaluation*, *19*(2–3), 170–187.

Sjøberg, S. (2015). PISA and global educational governance – A critique of the project, its uses and implications. *Eurasia Journal of Mathematics, Science and Technology Education*, *11*(1), 111–127. https://doi.org/10.12973/eurasia.2015.1310a

Smith, E. (2008). Using secondary data in educational and social research. McGraw-Hill Education.

Snow, C. E. (2002). Reading for understanding: Toward an R&D program in reading comprehension. Rand Corporation.

Snow, C. E. (2018). Simple and not-so-simple views of reading. *Remedial and Special Education*, *39*(5), 313–316. https://doi.org/10.1177/0741932518770288

Snowling, M. J., Lervåg, A., Nash, H. M., & Hulme, C. (2019). Longitudinal relationships between speech perception, phonological skills and reading in children at high-risk of dyslexia. *Developmental Science*, *22*(1), e12723. https://doi.org/10.1111/desc.12723

Soares De Sousa, D. (2011). Learning to read in English: Comparing monolingual English and bilingual Zulu-English Grade 3 learners. *South African Journal of Childhood Education*, *1*(1), 1–18. https://doi.org/10.4102/sajce.v1i1.70

South African Schools, Act No. 84 of 1996.

Spaull, N. (2013). Poverty & privilege: Primary school inequality in South Africa. *International Journal of Educational Development*, *33*(5), 436–447. https://doi.org/10.1016/j.ijedudev.2012.09.009

Spaull, N. (2015, June). The limitations of the Annual National Assessments (ANAs): What can and can't we say?. Presentation to the CEPD-Umalusi Wits School of Education.

Spaull, N. (2019). Equity: A price too high to pay? In N. Spaull & J. D. Jansen (Eds.), *South African schooling: The enigma of inequality* (pp. 1–24). Springer International Publishing.

Spaull, N. (2022). 2022 background report for the 2030 reading panel.

Spaull, N., & Jansen, J. D. (Eds.). (2019). *South African schooling: The enigma of inequality*. Springer.

Spaull, N., & Pretorius, E. J. (2019). Still falling at the first hurdle: Examining early grade reading in South Africa. In N. Spaull & J. D. Jansen (Eds.), *South African schooling: The enigma of inequality* (pp. 147–168). Springer.

Spaull, N., & Pretorius, E. J. (2022). Coming or going? The prioritisation of early grade reading in South Africa. In N. Spaull & E. Pretorius (Eds.), *Early grade reading in South Africa* (pp. 253–272). Oxford University Press.

Spaull, N., Pretorius, E. J., & Mohohlwane, N. L. (2020). Investigating the comprehension iceberg: Developing empirical benchmarks for early-grade reading in agglutinating African languages.

South African Journal of Childhood Education, *10*(1), 7–73. https://doi.org/10.4102/sajce.v10i1.773

Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly*, *16*(1), 32. https://doi.org/10.2307/747348

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods*, *11*(4), 402.

Steiner-Khamsi, G., & Waldow, F. (2018). PISA for scandalisation, PISA for projection: the use of international large-scale assessments in education policy making – an introduction. *Globalisation, Societies and Education*, *16*(5), 557–565. https://doi.org/10.1080/14767724.2018.1531234

Stiff, J. C., Lenkeit, J., Hopfenbeck, T. N., Kayton, H. L., & McGrane, J. A. (2023). Research engagement in the Progress in International Reading Literacy Study: A systematic review. *Educational Research Review*, *40*, 100547. https://doi.org/10.1016/j.edurev.2023.100547

Stoet, G., & Geary, D. C. (2013). Sex differences in mathematics and reading achievement are inversely related: Within- and across-nation assessment of 10 years of PISA data. *Plos One*, *8*(3), e57988. https://doi.org/10.1371/journal.pone.0057988

Stoffelsma, L. (2019). English vocabulary exposure in South African township schools: Pitfalls and opportunities. *Reading & Writing – Journal of the Reading Association of South Africa*, *10*(1), 1–10.

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.

Svetina, D., Crawford, A. V, Levy, R., Green, S. B., Scott, L., Thompson, M., Gorin, J. S., Fay, D., & Kunze, K. L. (2013). Designing small-scale tests: A simulation study of parameter recovery with the 1-PL. *Psychological Test and Assessment Modeling*, *55*(4), 335–360.

Svetina Valdivia, D., Rutkowski, L., Rutkowski, D., Canbolat, Y., & Underhill, S. (2023). Test engagement and rapid guessing: Evidence from a large-scale state assessment. *Frontiers in Education*, *8*, 1127644. https://doi.org/10.3389/feduc.2023.1127644

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). *Using multivariate statistics* (Vol. 5). Pearson.

Taboada, A., Tonks, S. M., Wigfield, A., & Guthrie, J. T. (2009). Effects of motivational and cognitive variables on reading comprehension. *Reading and Writing*, *22*(1), 85.

Tay, L., Meade, A. W., & Cao, M. (2014). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods*, *18*(1), 3–46. https://doi.org/10.1177/1094428114553062

Taylor, L. K., Bernhard, J. K., Garg, S., & Cummins, J. (2008). Affirming plural belonging: Building on students' family-based cultural and linguistic capital through multiliteracies pedagogy. *Journal of Early Childhood Literacy*, *8*(3), 269–294. https://doi.org/10.1177/1468798408096481

Taylor, N. (2023). The PIRLS assessment framework: Implications for teacher education. In S. van Staden & C. Combrinck (Eds.), *Tracking changes in South African Reading literacy achievement: A developing context perspective* (pp. 17–35). Brill.

Taylor, N., & Taylor, S. (2013). Learner performance in the National Schools Effectiveness Study. In N. Taylor, S. van der Berg, & T. Mabogoane (Eds.), *Creating effective schools: Report of South Africa's National Schools Effectiveness Study*. Pearson.

Taylor, S. (2017). Improved reading in early grades will transform South Africa. *Business Day*.

Taylor, S. (2019). How can learning inequalities be reduced? Lessons learnt from experimental research in South Africa. In N. Spaull & J. D. Jansen (Eds.), *South African schooling: The enigma of inequality* (pp. 321–336). Springer. https://doi.org/10.1007/978-3-030-18811-5_17

Taylor, S., & von Fintel, M. (2016). Estimating the impact of language of instruction in South African primary schools: A fixed effects approach. *Economics of Education Review*, *50*, 75–89. https://doi.org/10.1016/j.econedurev.2016.01.003

Taylor, S., & Yu, D. (2009). *The importance of socio-economic status in determining educational achievement in South Africa* (Stellenbosch Economic Working Papers No. 01). http://hdl.handle.net/10019.1/41592

Temel, G. Y. (2023). A simulation and empirical study of differential test functioning (DTF). *Psych*, *5*(2), 478–496. https://doi.org/10.3390/psych5020032

Tenopyr, M. L. (1977). Content-construct confusion. *Personnel Psychology*, *30*(1), 47–54. https://doi.org/10.1111/j.1744-6570.1977.tb02320.x

Teresi, J. A., Wang, C., Kleinman, M., Jones, R. N., & Weiss, D. J. (2021). Differential item functioning analyses of the Patient-Reported Outcomes Measurement Information System (PROMIS®) measures: Methods, challenges, advances, and future directions. *Psychometrika*, *86*(3), 674–711. https://doi.org/10.1007/S11336-021-09775-0

Thissen, D. (2001). IRTLRDIF v.2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning. L.L. Thurstone Psychometric Laboratory.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 63–113). Lawrence Erlbaum Associates.

Thompson, C. G., Kim, R. S., Aloe, A. M., & Becker, B. J. (2017). Extracting the variance inflation factor and other multicollinearity diagnostics from typical regression results. *Basic and Applied Social Psychology*, *39*(2), 81–90. https://doi.org/10.1080/01973533.2016.1277529

Thomson, S. (2021). PISA Australia – Excellence and equity? In N. Crato (Ed.), *Improving a country's education: PISA 2018 results in 10 countries* (pp. 25–46). Springer.

Thorndike, E. L. (1917). Reading as reasoning: A study of mistakes in paragraph reading. *Journal of Educational Psychology*, *8*(6), 323–330. https://doi.org/10.1037/h0075325

Tijmstra, J., Bolsinova, M., Liaw, Y.-L., Rutkowski, L., & Rutkowski, D. (2020). Sensitivity of the RMSD for detecting item-level misfit in low-performing countries. *Journal of Educational Measurement*, *2020*(4), 566–583. https://doi.org/10.1111/jedm.12263

Toste, J. R., Didion, L., Peng, P., Filderman, M. J., & McClelland, A. M. (2020). A meta-analytic review of the relations between motivation and reading achievement for K-12 students. *Review of Educational Research*, *90*(3), 420–456.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement*, *16*, 8–13.

Trudell, B., & Schroeder, L. (2007). Reading methodologies for African languages: Avoiding linguistic and pedagogical imperialism. *Language, Culture and Curriculum*, *20*(3), 165–180. https://doi.org/10.2167/lcc333.0

UNESCO. (1962). Educational achievements of thirteen-year-olds in twelve countries: Results of an international research project, 1959-61. UNESCO Institute for Education. https://unesdoc.unesco.org/ark:/48223/pf0000131437

UNESCO. (2014). Education for all global monitoring report 2013/4: Teaching and learning (GEM Report). UNESCO GEM Team. https://doi.org/10.54676/CMSE2898

UNESCO. (2015). Education for all global monitoring report. Regional overview: Sub-Saharan Africa. UNESCO.

UNESCO, & IEA. (2017). Measuring SDG 4: How PIRLS can help. International Association for the Evaluation of Educational Achievement.

United Nations. (2015). Sustainable Development Goal 4: Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all. United Nations Sustainable Development Goals. https://sdgs.un.org/goals/goal4

United Nations. (2020). Human development report 2020.

Uys, M., van der Walt, J., van den Berg, R., & Botha, S. (2007). English medium of instruction: A situation analysis. *South African Journal of Education*, *27*(1), 69–82.

Vainikainen, M. P., Hienonen, N., & Hotulainen, R. (2017). Class size as a means of three-tiered support in Finnish primary schools. *Learning and Individual Differences*, *56*, 96–104. https://doi.org/10.1016/j.lindif.2017.05.004

van Bergen, E., van Zuijen, T., Bishop, D., & de Jong, P. F. (2017). Why are home literacy environment and children's reading skills associated? What parental skills reveal. *Reading Research Quarterly*, *52*(2), 147–160. https://doi.org/10.1002/RRQ.160

van de Vijver, F. J. R., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Revue Europeenne de Psychologie Appliquee*. https://doi.org/10.1016/j.erap.2003.12.004

van der Berg, S. (2007). Apartheid's enduring legacy: Inequalities in education. *Journal of African Economies*, *16*(5), 849–880.

van der Berg, S. (2015). What the Annual National Assessments can tell us about learning deficits over the education system and the school career. *South African Journal of Childhood Education*, *5*(2), 16. https://doi.org/10.4102/sajce.v5i2.389

van der Berg, S., Burger, C., Burger, R., de Vos, M., du Rand, G., Gustafsson, M., Moses, E., Shepherd, D. L., Spaull, N., Taylor, S., van Broekhuizen, H., & von Fintel, D. (2011). Low quality education as a poverty trap (Stellenbosch Economic Working Papers No, 25). Synthesis report to PSPPD. https://doi.org/10.2139/ssrn.2973766

van der Berg, S., & Gustafsson, M. (2019). Educational outcomes in post-apartheid South Africa: Signs of progress despite great inequality. In N. Spaull & J. Jansen (Eds.), *South African schooling: The enigma of inequality* (pp. 25–46). Springer International Publishing.

van der Berg, S., Hoadley, U., Galant, J., van Wyk, C., & Böhmer, B. (2022). *Learning losses from COVID-19 in the Western Cape: Evidence from Systemic Tests*. ReSEP, Stellenbosch University.

van der Berg, S., Spaull, N., Wills, G., Gustafsson, M., & Kotzé, J. (2016). Identifying binding constraints in education: Synthesis report for the programme to support pro-poor policy development (PSPPD). http://dx.doi.org/10.2139/ssrn.4212977 van Hek, M., Kraaykamp, G., & Wolbers, M. H. J. (2016). Comparing the gender gap in educational attainment: The impact of emancipatory contexts in 33 cohorts across 33 countries. *Educational Research and Evaluation*, *22*(5–6), 260–282. https://doi.org/10.1080/13803611.2016.1256222

van Rijn, P. W., Sinharay, S., Haberman, S. J., & Johnson, M. S. (2016). Assessment of fit of item response theory models used in large-scale educational survey assessments. *Large-Scale Assessments in Education*, *4*(1). https://doi.org/10.1186/S40536-016-0025-3

van Rooy, B., & Pretorius, E. J. (2013). Is reading in an agglutinating language different from an analytic language? An analysis of isiZulu and English reading based on eye movements. *Southern African Linguistics and Applied Language Studies*, *31*(3), 281–297.

van Staden, S., & Combrinck, C. (Eds.). (2023). *Tracking changes in South African Reading literacy achievement: A developing context perspective*. Brill.

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), a712. https://doi.org/10.4102/sajce.v9i1.712

van Staden, S., & Gustafsson, M. (2022). What a decade of PIRLS results reveals about early grade reading in South Africa: 2006, 2011, 2016. In N. Spaull & E. Pretorius (Eds.), *Early grade reading in South Africa* (pp. 22–36). Oxford University Press.

van Staden, S., & Howie, S. (2012). Reading between the lines: Contributing factors that affect Grade 5 student reading performance as measured across South Africa's 11 languages. *Educational Research and Evaluation*, *18*(1), 85–98. https://doi.org/10.1080/13803611.2011.641270

van Staden, S., & Howie, S. (2014). Reflections on Creemers' comprehensive model of educational effectiveness for reading literacy: South African evidence from PIRLS 2006. *Perspectives in Education, 32*(3), 172-192.

Vaughn, S., Capin, P., Scammacca, N., Roberts, G., Cirino, P., & Fletcher, J. M. (2020). The critical role of word reading as a predictor of response to intervention. *Journal of Learning Disabilities*, *53*(6), 415–427. https://doi.org/10.1177/0022219419891412

Verger, A., Parcerisa, L., & Fontdevila, C. (2019). The growth and spread of large-scale assessments and test-based accountabilities: a political sociology of global education reforms. *Educational Review*, *71*(1), 5–30. https://doi.org/10.1080/00131911.2019.1522045

Vermeer, A. (2004). The relation between lexical richness and vocabulary size in Dutch L1 and L2 children. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language* (Vol. 10, pp. 173–189). https://doi.org/10.1075/lllt.10.13ver

Volante, L. (Ed.). (2017). *The PISA Effect on Global Educational Governance*. Routledge.

Volante, L., & Klinger, D. A. (2022). PISA, global reference societies, and policy borrowing: The promises and pitfalls of 'academic resilience.' *Policy Futures in Education*, *21*(7), 755–764. https://doi.org/10.1177/14782103211069002

Volante, L., Schnepf, S. V, & Klinger, D. A. (2022). *Cross-national achievement surveys for monitoring educational outcomes: Policies, practices and political reforms within the European Union.* Publications Office of the European Union. https://data.europa.eu/doi/10.2760/406165

von Davier, M., Gonzalez, E., & Schulz, W. (2020). Ensuring validity in international comparisons using state-of-the-art psychometric methodologies. In H. Wagemaker (Ed.), *Reliability and validity of international large-scale assessment: Understanding IEA's comparative studies of student achievement* (Vol. 10, pp. 187–220). Springer.

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. https://pirls2021.org/methods/

von Davier, M., & Sinharay, S. (2014). Analytics in international large-scale assessments: Item response theory and population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 155–174). CRC Press.

von Davier, M., Yamamoto, K., Shin, H. J., Chen, H., Khorramdel, L., Weeks, J., Davis, S., Kong, N., & Kandathil, M. (2019). Evaluating item response theory linking and model fit for data from PISA 2000-2012. *Assessment in Education: Principles, Policy & Practice*, *26*(4), 466–488. https://doi.org/10.1080/0969594X.2019.1586642

Wagemaker, H. (2020). Introduction to reliability and validity of international large-scale assessment. In H. Wagemaker (Ed.), Reliability and validity of international large-scale assessment: Understanding IEA's comparative studies of student achievement (Vol. 10, pp. 1–5). Springer. https://doi.org/10.1007/978-3-030-53081-5_1

Wagner, D. A. (2010). Quality of education, comparability, and assessment choice in developing countries. *Compare: A Journal of Comparative and International Education*, *40*(6), 741–760. https://doi.org/10.1080/03057925.2010.523231

Wagner, D. A. (2017). Children's reading in low-income countries. *Reading Teacher*, *71*(2), 127–133. https://doi.org/10.1002/trtr.1621

Wagner, D. A., Babson, A., & Murphy, K. (2011). How much is learning measurement worth? Assessment costs in low-income countries. *Current Issues in Comparative Education*, *14*(1).

Wagner, D. A., & Castillo, N. M. (2014). Learning at the bottom of the pyramid: Constraints, comparability and policy in developing countries. *Prospects*, *44*(4), 627–638. https://doi.org/10.1007/S11125-014-9328-8

Wainer, H. (1993). Model-based standardized measurement of an item's differential impact. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 123–135). Lawrence Erlbaum Associates.

Waldow, F. (2009). What PISA did and did not do: Germany after the "PISA-shock." *European Educational Research Journal*, *8*(3), 476–483. https://doi.org/10.2304/EERJ.2009.8.3.476

Wang, J., & Wu, Y. (2023). Income inequality, cultural capital, and high school students' academic achievement in OECD countries: A moderated mediation analysis. *The British Journal of Sociology*, *74*(2), 148–172. https://doi.org/10.1111/1468-4446.12997

Wang, M., & Woods, C. M. (2016). Anchor selection using the Wald test anchor-all-test-all procedure. *Applied Psychological Measurement*, *41*(1), 17–29. https://doi.org/10.1177/0146621616668014

Wang, Z. (2020). When large-scale assessments meet data science: The big-fish-little-pond effect in fourth- and eighth-grade mathematics across nations. *Frontiers in Psychology*, *11*, 579545. https://doi.org/10.3389/fpsyg.2020.579545

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples and criteria. *Journal of Applied Psychology*, *52*(5), 372–376. https://doi.org/10.1037/H0026244

Westerveld, M. F., Armstrong, R. M., & Barton, G. M. (2020). *Reading success in the primary years: An evidence-based interdisciplinary approach to guide assessment and Intervention*. Springer Nature. https://doi.org/10.1007/978-981-15-3492-8

Wigfield, A., Gladstone, J. R., & Turci, L. (2016). Beyond cognition: Reading motivation and reading comprehension. *Child Development Perspectives*, *10*(3), 190–195. https://doi.org/10.1111/cdep.12184

Wilsenach, C. (2013). Phonological skills as predictor of reading success: An investigation of emergent bilingual Northern Sotho/English learners. *Per Linguam: A Journal of Language Learning, Per Linguam: Tydskrif vir Taalaanleer*, *29*(2), 17–32.

Wilsenach, C. (2015). Receptive vocabulary and early literacy skills in emergent bilingual Northern Sotho-English children. *Reading & Writing – Journal of the Reading Association of South Africa*, *6*(1), 1–11.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching*, *46*(6), 716–730. https://doi.org/10.1002/tea.20318

Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, *15*(1), 27–41. https://doi.org/10.1080/10627191003673216

Wise, S. L., Soland, J., & Bo, Y. (2020). The (non)impact of differential test taker engagement on aggregated scores. *International Journal of Testing*, *20*(1), 57–77. https://doi.org/10.1080/15305058.2019.1605999

Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequence of performance, test, motivation, and mentally taxing items. *Applied Measurement in Education*, *8*(4), 341–351. https://doi.org/10.1207/S15324818AME0804_4

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock–Johnson III tests of achievement*. Riverside.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA press.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Mesa press.

Wright, K. D., & Oshima, T. C. (2015). An effect size measure for Raju's differential functioning for items and tests. *Educational and Psychological Measurement*, *75*(2), 338–358. https://doi.org/10.1177/0013164414532944

Yamamoto, K., Khorramdel, L., & Shin, H. J. (2018). Introducing multistage adaptive testing into international large-scale assessments designs using the example of PIAAC. *Psychological Test and Assessment Modeling*, *60*(3), 347–368.

Yamamoto, K., Shin, H. J., & Khorramdel, L. (2018). Multistage adaptive testing design in international large-scale assessments. *Educational Measurement: Issues and Practice*, *37*(4), 16–27. https://doi.org/10.1111/emip.12226

Yamashita, J. (2002). Influence of L1 reading on L2 reading: Different perspectives from the process and product of reading. *Studies in Language and Culture*, *23*(2), 271–283.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125–145. https://doi.org/10.1177/014662168400800201

Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology*, *9*(2), 79–94.

Yoshino, A. (2012). The relationship between self-concept and achievement in TIMSS 2007: A comparison between American and Japanese students. *International Review of Education*, *58*(2), 199–219. https://doi.org/10.1007/S11159-012-9283-7

Zhao, X. (2018). *Test translation review procedures in international large-scale assessment: Sensitivity to culture and society* [Doctoral dissertation, University of Colorado]. ProQuest Dissertations Publishing.

Zhao, Y. (2016). Who's afraid of PISA: The fallacy of international assessments of system performance. In A. Harris & M. S. Jones (Eds.), *Leading futures: Global perspectives on educational leadership* (pp. 7–22). SAGE.

Zhao, Y. (2020). Two decades of havoc: A synthesis of criticism against PISA. *Journal of Educational Change*, *21*(2), 245–266. https://doi.org/10.1007/S10833-019-09367-X

Zhao, Y., & Hambleton, R. K. (2017). Practical consequences of item response theory model misfit in the context of test equating with mixed-format test data. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00484

Ziglari, L. (2017). Interpreting multiple regression results: β weights and structure coefficients. *General Linear Model Journal*, *43*(2), 13–22.

Zimmerman, L. (2014). Lessons learnt: Observation of Grade 4 reading comprehension teaching in South African schools across the Progress in International Reading Literacy Study (PIRLS) 2006 achievement spectrum. *Reading & Writing – Journal of the Reading Association of South Africa*, *5*(1), e1–e9. https://doi.org/10.4102/rw.v5i1.48

Zimmerman, L., & Smit, B. (2014). Profiling classroom reading comprehension development practices from the PIRLS 2006 in South Africa. *South African Journal of Education*, *34*(3), 1–9. https://doi.org/10.15700/201409161101

Zimmerman, L., & Smit, B. (2016). Comprehending the macro through the lens of the micro: The use of PIRLS 2006 findings to inform comparative case studies across the South African achievement scale. *International Journal of Qualitative Methods*, *15*(1). https://doi.org/10.1177/1609406915624576

Zulu, P. N., Botha, G., & Barnard, E. (2008). Orthographic measures of language distances between the official South African languages. *Literator: Journal of Literary Criticism, Comparative Linguistics and Literary Studies*, *29*(1), 185–204.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*(2), 223–233. https://doi.org/10.1080/15434300701375832

Zuze, T. L., & Hannan, S. (2016). *The story of assessments in South Africa*. TIMSS SA Newsletter. https://www.timss-sa.org/blog/the-story-of-assessments-in-south-africa (Accessed 20 September 2023)

APPENDICES

Appendix A PIRLS 2016 instrument

PIRLS 2016 consists of three distinct tests: PIRLS 2016, PIRLS Literacy 2016, and ePIRLS 2016. PIRLS 2016 is the main test of reading achievement; it consists of 12 reading passages and approximately 190 corresponding items. Each of the passages is approximately 400-800 words long. Each student is presented with only two passages and each passage has approximately 15 corresponding items. PIRLS Literacy 2016 (previously prePIRLS) is designed to be a less difficult, but still scalable, version of the PIRLS assessment. PIRLS Literacy 2016 was primarily used by participating countries as a benchmarking instrument. ePIRLS 2016 is a test of online reading first implemented in the 2016 cycle, it is designed to incorporate the changing nature of reading into the PIRLS assessment. This section provides additional detail regarding the PIRLS Literacy 2016 instrument, including example passages and items, as well as a description of the scoring and benchmarking procedures for PIRLS 2016.

A-1 Description of PIRLS Literacy 2016 passages and items

The passages that form part of PIRLS Literacy 2016 are separated into two categories: literary texts, and informational texts. The literary texts consist of contemporary short stories, traditional tales, and fables. The informational texts come from a variety of sources and reflect textbook or article formats with information presented in text as well as in graphics such as diagrams and charts. The passages cover a range of reading material that is selected from submissions by participating countries to represent the authentic reading experiences of students both in school and in their everyday lives (Martin, Mullis, & Foy, 2017).

Before being selected to form part of the instrument, each passage underwent qualitative evaluation by experts at the IEA as well as by National Research Coordinators (NRCs) from each country to assess the passages in terms of the text length, characteristics, readability, and appropriateness for translation into different languages and cultures (Martin, Mullis, & Hooper, 2017). Additionally, quantitative evaluation was provided to NRCs for each passage that detailed the word count and the Flesch-Kincaid Grade Level Formula as an indication of readability for each text. Table A-1 provides the titles and a brief description of each of the 12 selected texts that form part of dataset as well as their word count in English, a Flesch-Kinkaid readability index in English, and the number of items per passage (Martin, Mullis, & Foy, 2017).

Table A-1

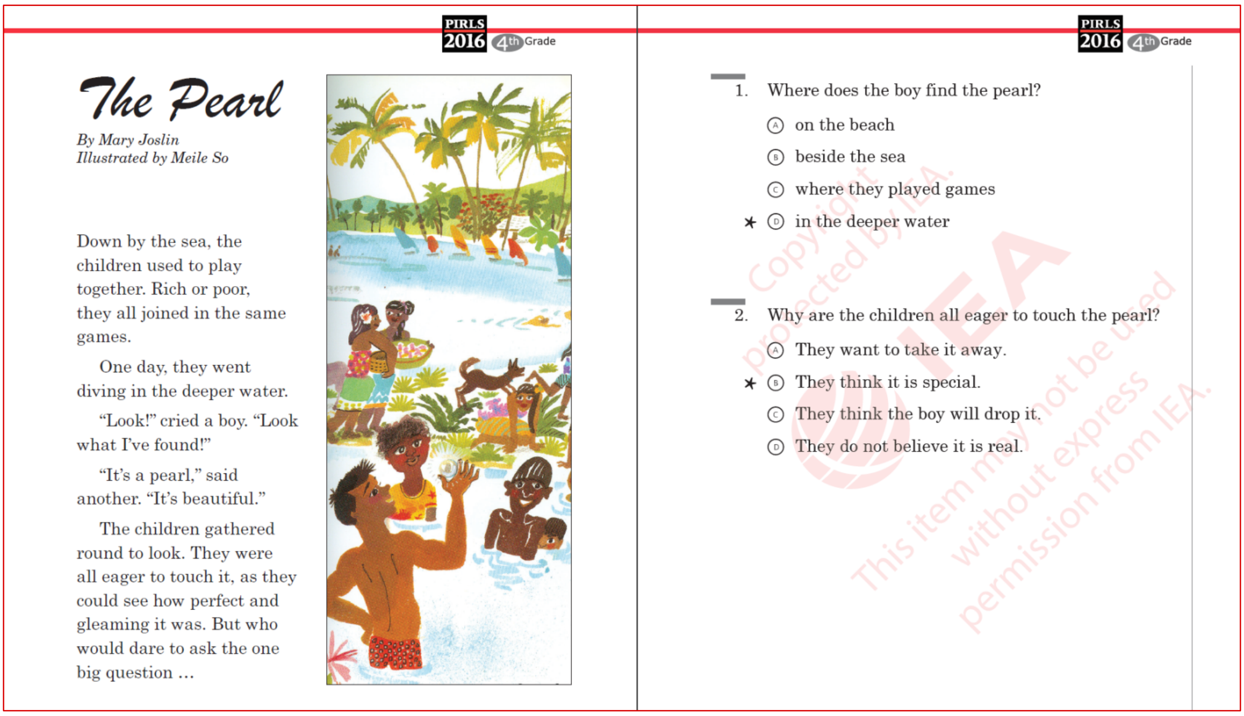
PIRLS Literacy 2016 passage titles and description

Passage title	Details	Word count	Flesch-Kincaid	Number of items
	Reading for literary experience			
Flowers on the roof†	This contemporary story portrays friendship between the generations.	811	2.8	13
Pemba Sherpa†	This modern tale set in the Himalayan Mountains tells the story of a young girl determined to be a sherpa.	540	2.5	17
Baghita's Perfect Orange	This traditional tale set in Africa has a moral about greed and generosity.	404	2.0	16
The Pearl	This story about a young pearl merchant illustrates the power of home, friendship, and generosity above greed.	536	2.9	15
The Summer My Father Was Ten	In this thought-provoking story with a realistic contemporary setting, a boy is allowed to make amends for his thoughtless behavior.	484	4.0	14
Library Mouse	This story is about a mouse who lives in the library and inspires young children to be authors.	497	3.1	18
	Reading to acquire and use information			
How Did We Learn to Fly?†	This historical text explains how the modern airplane was developed.	514	6.3	17
Sharks†	This article presents information about sharks in a variety of formats, using subheadings, a labeled diagram, and photographs.	570	7.6	12
Hungry Plant	This scientific text describes the Venus Flytrap plant and explains how it captures insects for food.	509	3.5	16
Ants	This article presents information about the lives of different types of ants, using subheadings, photographs, and diagrams.	415	2.9	14

Passage title	Details	Word count	Flesch-Kincaid	Number of items
African Rhinos & Oxpecker Birds	This passage presents information about African rhinos and oxpecker birds and describes how the two animals depend on one another for food and survival.	449	4.7	17
Training a Deaf Polar Bear	The passage describes how zookeepers worked with a polar bear that was found to be deaf.	425	4.0	14

Source: PIRLS 2016 Assessment Framework
Note: † indicates passages shared with main PIRLS test

As is shown in the table, the passages are between 400 and 600 words long, with one longer passage of just over 800 words. There are approximate 15 corresponding items for each passage which are interspersed within the passage allowing the students to respond as they are reading. In other words, a small group of items (~3) follow a small piece of the text (~80 words) – this serves to support students' retrieval of information by making it clear which section of the passage is relevant for those items. Figure A-1 provides an example of one of the items from a PIRLS Literacy 2016 passage and its corresponding items.


PIRLS 2016 4th Grade

The Pearl

By Mary Joslin
Illustrated by Meile So

Down by the sea, the children used to play together. Rich or poor, they all joined in the same games.

One day, they went diving in the deeper water.

"Look!" cried a boy. "Look what I've found!"

"It's a pearl," said another. "It's beautiful."

The children gathered round to look. They were all eager to touch it, as they could see how perfect and gleaming it was. But who would dare to ask the one big question …

PIRLS 2016 4th Grade

1. Where does the boy find the pearl?
- Ⓐ on the beach
- Ⓑ beside the sea
- Ⓒ where they played games
- ✶ Ⓓ in the deeper water

2. Why are the children all eager to touch the pearl?
- Ⓐ They want to take it away.
- ✶ Ⓑ They think it is special.
- Ⓒ They think the boy will drop it.
- Ⓓ They do not believe it is real.

Source: PIRLS 2016 Assessment Framework

Figure A-1 Example excerpt and items from "The Pearl" by Mary Joslin

An important consideration in the assessment procedures outlined by PIRLS Literacy 2016 is the time students take to complete the assessment. To limit the assessment burden, each student answers items based on only two of the passages which are distributed into booklets with an expected testing time of 40 minutes for students.

A-2 Booklet Design

The design of the booklets for PIRLS Literacy 2016 followed a similar procedure used in the main PIRLS 2016 assessment. Each student who took part in PIRLS Literacy 2016 receives one booklet that contains two passages and the corresponding items for those passages. To ensure an even distribution of passages to students, the passages are organised into blocks. Five literary blocks which contain passages and items related to reading for literary experience, and five informational blocks which contain passages and items related to reading to acquire and use information were created. These blocks were then distributed into 15 booklets using a matrix sampling technique where each passage and its items are assigned to a block and these blocks are then systematically distributed into the student booklets. This means that each passage occurs three times across the 15 booklets, in different pairs. As passages in PIRLS are more difficult than those in PIRLS Literacy (by design), when passages shared with PIRLS are included in the booklets, they always appear in the second position. The 16th booklet, referred to as the *Reader*, contains a literary and informational passage that do not overlap with the other booklets but are shared across PIRLS and PIRLS Literacy 2016. As a result of this, the Reader is assigned three times more frequently in the randomised distribution procedure than the other booklets to ensure that the proportion of students responding to the passages in the Reader is the same as the proportion of students responding to the other passages. Table A-2 provides an illustration of the matrix design applied to distribute the 12 passages across all 16 booklets.

Table A-2

Matrix design for passage distribution across booklets

Booklet	**Part 1**	**Part 2**
Booklet 1	Hungry plant	Flowers on the roof
Booklet 2	The pearl	Hungry plant
Booklet 3	African rhinos & oxpecker birds	The pearl
Booklet 4	African rhinos & oxpecker birds	Library mouse
Booklet 5	Library mouse	How did we learn to fly?
Booklet 6	Bhaghita's perfect orange	How did we learn to fly?
Booklet 7	Ants	Bhaghita's perfect orange
Booklet 8	The summer my father was ten	Ants
Booklet 9	The summer my father was ten	Training a deaf polar bear
Booklet 10	Training a deaf polar bear	Flowers on the roof
Booklet 11	The pearl	How did we learn to fly?
Booklet 12	African rhinos & oxpecker birds	Bhaghita's perfect orange
Booklet 13	Ants	Flowers on the roof
Booklet 14	The summer my father was ten	Hungry plant
Booklet 15	Training a deaf polar bear	Library mouse
Reader	Pemba sherpa	Sharks

Source: PIRLS 2016 Assessment Framework

A-3 Scoring

PIRLS Literacy 2016 uses both multiple-choice and constructed response items to assess the reading ability of students. The items that correspond with each passage are designed to assess each of the four processes of comprehension outlined in the reading framework in both the literary and informational texts (Martin, Mullis, & Hooper, 2017). Each item has a focus on a specific comprehension process, for instance, an item aimed to assess ‘making straightforward inferences’ would require students to demonstrate their ability to infer the connection between two or more pieces of information provided in the passage. Multiple-choice items require students to select the correct answer from presented choices and items have either two (A, B) or four (A, B, C, D) possible answer choices. The 94 multiple-choice items are all dichotomously scored (0 – incorrect or 1 – correct). Constructed response items require students to write their own answer on the line provided for the item. The 96 constructed response items are scored in three possible categories; 69 are scored dichotomously (0 – incorrect or 1 – correct); 23 are scored polytomously with a maximum of 2 possible points (0 – incorrect, 1 – partially correct, 2 – correct); and 4 are scored polytomously with a maximum of 3 possible points (0 – incorrect, 1 – partially correct, 2 – almost correct; 3 – correct). Table A-3 provides a summary of the number of items for each item type, scoring type and maximum score included in the dataset.

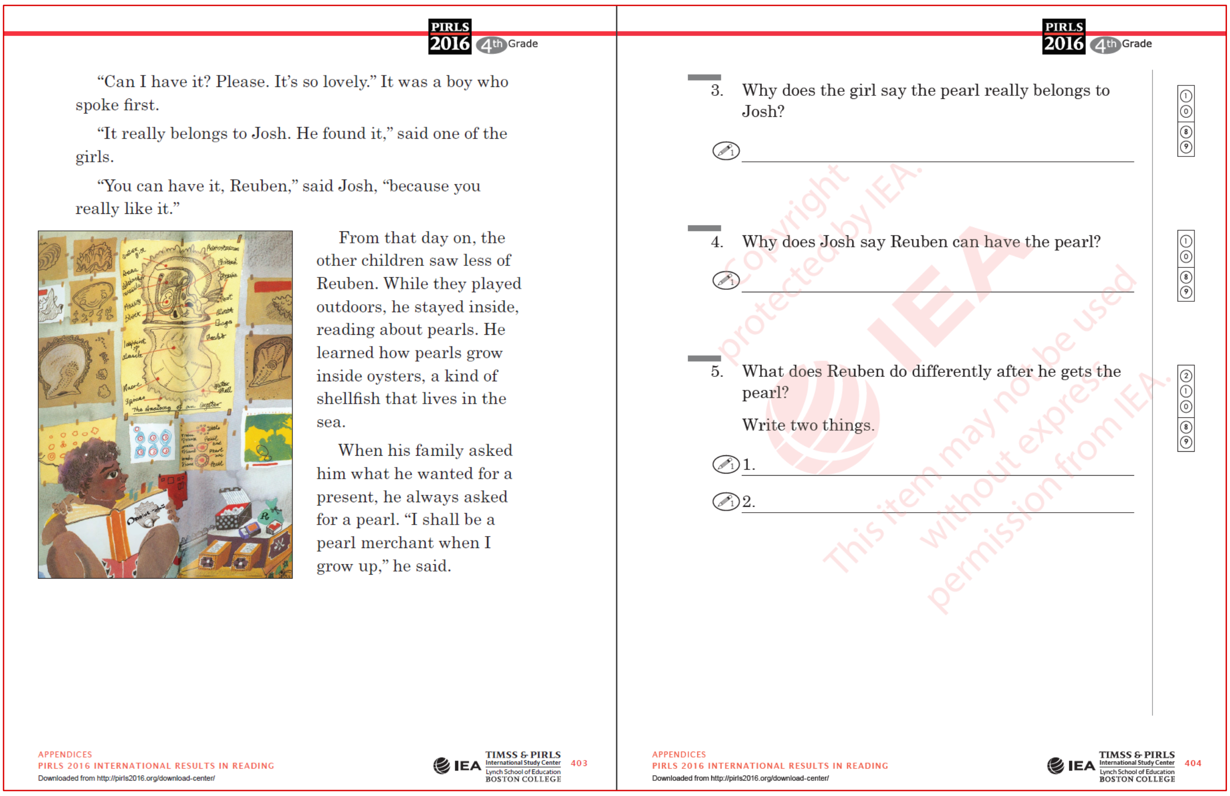
Table A-3

Item types, scoring procedures and models applied in PIRLS Literacy 2016

Item type	Scoring type	Maximum score	Scoring	Number of items
Multiple-choice	Dichotomous	1	0 – incorrect 1 – correct	94
Constructed response	Dichotomous	1	0 – incorrect 1 – correct	69
	Polytomous	2	0 – incorrect 1 – partially correct 2 – correct	23
		3	0 – incorrect 1 – partially correct 2 – almost correct 3 – correct	4
			Total	190

To ensure scoring consistency, intensive training is provided to NRCs and their staff, and a scoring guide is provided that details the options and criteria for awarding different scores for each item (Martin, Mullis, & Hooper, 2017). These scoring guides provide guidelines for scorers in participating countries to follow when making decisions about how to award different student responses to constructed-response items. The focus of each item, the essential features of appropriate responses, as well as the criteria for determining complete responses are described.

The guidelines for the scoring of PIRLS Literacy 2016 explain that the focus is on the students' ability to demonstrate that they have understood the text, and not on perfectly written responses. This means that single-word answers, misspelled answers and answers with incorrect grammar can still be awarded full scores if they fall within the provided guidelines and demonstrate understanding of the passage and item. Figure A-2 and Figure A-3 provide an example of two constructed response items as well as the scoring guide for each from the passage "The Pearl".



PIRLS 2016 4th Grade

"Can I have it? Please. It's so lovely." It was a boy who spoke first.

"It really belongs to Josh. He found it," said one of the girls.

"You can have it, Reuben," said Josh, "because you really like it."

From that day on, the other children saw less of Reuben. While they played outdoors, he stayed inside, reading about pearls. He learned how pearls grow inside oysters, a kind of shellfish that lives in the sea.

When his family asked him what he wanted for a present, he always asked for a pearl. "I shall be a pearl merchant when I grow up," he said.

APPENDICES
PIRLS 2016 INTERNATIONAL RESULTS IN READING
Downloaded from http://pirls2016.org/download-center/
IEA TIMSS & PIRLS International Study Center Lynch School of Education BOSTON COLLEGE 403

PIRLS 2016 4th Grade

3. Why does the girl say the pearl really belongs to Josh?

4. Why does Josh say Reuben can have the pearl?

5. What does Reuben do differently after he gets the pearl?

Write two things.

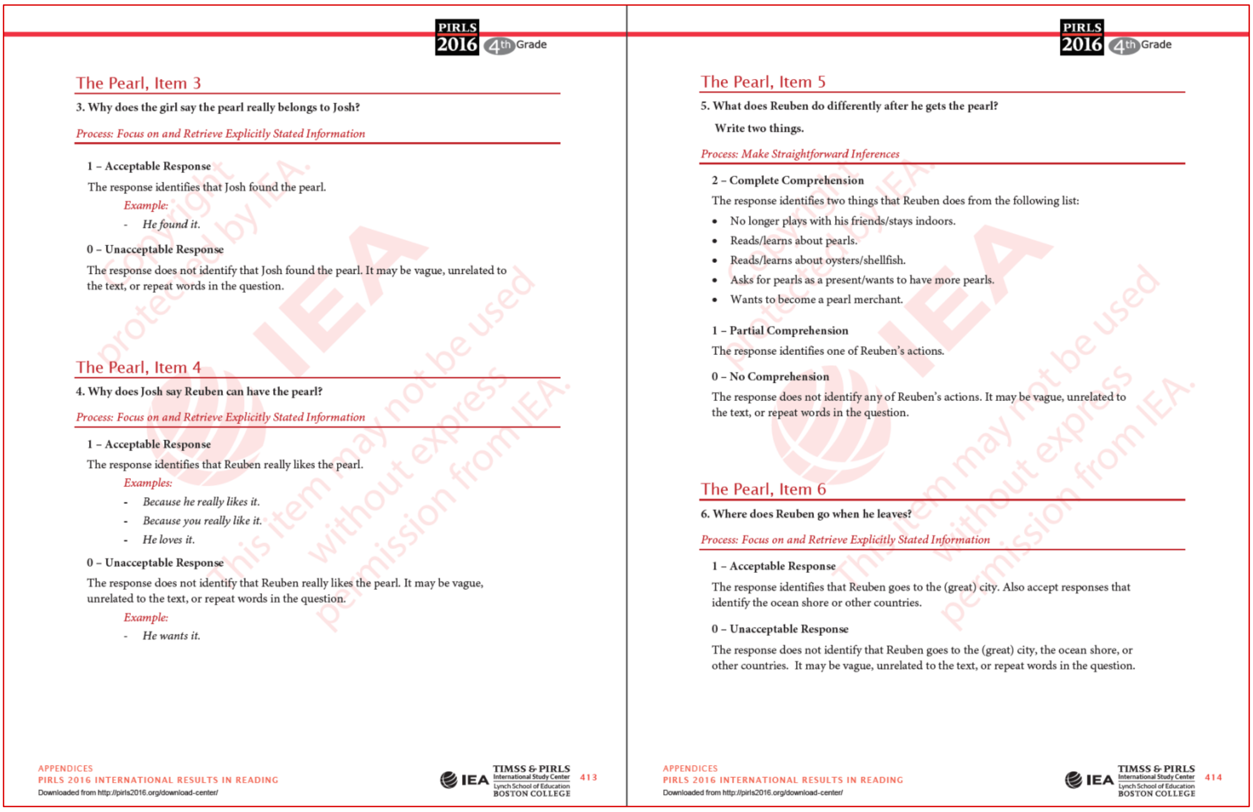
1.

2.

APPENDICES
PIRLS 2016 INTERNATIONAL RESULTS IN READING
Downloaded from http://pirls2016.org/download-center/
IEA TIMSS & PIRLS International Study Center Lynch School of Education BOSTON COLLEGE 404

Source: PIRLS 2016 Assessment Framework

Figure A-2 Example of constructed response items from "The Pearl"



PIRLS 2016 4th Grade

The Pearl, Item 3

3. Why does the girl say the pearl really belongs to Josh?

Process: Focus on and Retrieve Explicitly Stated Information

1 – Acceptable Response

The response identifies that Josh found the pearl.

Example:

- *He found it.*

0 – Unacceptable Response

The response does not identify that Josh found the pearl. It may be vague, unrelated to the text, or repeat words in the question.

The Pearl, Item 4

4. Why does Josh say Reuben can have the pearl?

Process: Focus on and Retrieve Explicitly Stated Information

1 – Acceptable Response

The response identifies that Reuben really likes the pearl.

Examples:

- *Because he really likes it.*
- *Because you really like it.*
- *He loves it.*

0 – Unacceptable Response

The response does not identify that Reuben really likes the pearl. It may be vague, unrelated to the text, or repeat words in the question.

Example:

- *He wants it.*

PIRLS 2016 4th Grade

The Pearl, Item 5

5. What does Reuben do differently after he gets the pearl?

Write two things.

Process: Make Straightforward Inferences

2 – Complete Comprehension

The response identifies two things that Reuben does from the following list:

- No longer plays with his friends/stays indoors.
- Reads/learns about pearls.
- Reads/learns about oysters/shellfish.
- Asks for pearls as a present/wants to have more pearls.
- Wants to become a pearl merchant.

1 – Partial Comprehension

The response identifies one of Reuben's actions.

0 – No Comprehension

The response does not identify any of Reuben's actions. It may be vague, unrelated to the text, or repeat words in the question.

The Pearl, Item 6

6. Where does Reuben go when he leaves?

Process: Focus on and Retrieve Explicitly Stated Information

1 – Acceptable Response

The response identifies that Reuben goes to the (great) city. Also accept responses that identify the ocean shore or other countries.

0 – Unacceptable Response

The response does not identify that Reuben goes to the (great) city, the ocean shore, or other countries. It may be vague, unrelated to the text, or repeat words in the question.

Source: PIRLS 2016 Assessment Framework

Figure A-3 Example of a scoring guide for constructed response items for "The Pearl"

As is evident from the scoring guide, examples of both acceptable and unacceptable responses are provided to assist with scoring consistency. In an added measure to ensure consistency of scoring, inter-scorer reliability is measured within each country, across countries and over time (i.e., across different cycles) (Martin, Mullis, & Hooper, 2017). To measure reliability of scoring within countries a subset of 200 student responses to all constructed response items are randomly selected by each participating country to be scored independently by two scorers. The degree of agreement between these independent scorers is then reported to IEA as a measure of reliability. If the inter-rater reliability is below 85%, scorers are required to retrain and then rescore items. The average inter-rater reliability across items for PIRLS Literacy 2016 in South Africa was 93% (Martin, Mullis, & Hooper, 2017).

A-4 PIRLS Benchmarks

PIRLS provides an overall score for reading achievement for each student along a linear scale with a mean of 500 points and a standard deviation of 100 points. In addition to the overall score, reading proficiency scores are also provided for the two purposes of reading, as well as for the four comprehension processes. In an effort to enable straightforward interpretation of the reading achievement scores provided to participating countries, PIRLS provides four benchmarks of reading achievement along the scale: low (400-474), intermediate (475-549), high (550-625), and advanced (>625) (Mullis et al., 2015). Each of these benchmarks corresponds with a set of criteria which describes what students at each benchmark are able to demonstrate when reading literary and informational texts.

Table A-4

Description of benchmark criteria for PIRLS 2016

International Benchmark Level	When reading literary texts students can:	When reading informational texts students can:
Low benchmark 400-474	Locate and retrieve explicitly stated information	Locate and retrieve 2-3 pieces of information in text Find information in text boxes, headings and figures
Intermediate benchmark 475-549	Retrieve and reproduce explicit information Make straight-forward inferences about character feelings, motivations Interpret obvious reasons and causes, give basic explanations	Locate and reproduce 2-3 pieces of information from text Use sub-headings, figures and text boxes to locate information Retrieve and reproduce explicit information
High benchmark 550-625	Identify significant events & actions Make inferences and explain relationships, give text-based support Identify significance of events, recognise language features (tone)	Locate relevant information within complex text or table Make inferences and logical connections to provide explanations Evaluate content and make generalisations
Advanced benchmark 625+	Integrate ideas and evidence across a text to appreciate overall themes	Distinguish and interpret complex information from different parts of text

International Benchmark Level	When reading literary texts students can:	When reading informational texts students can:
	Interpret story events & character actions, provide insights that are text based Interpret story events & character actions, provide insights that are text based	Integrate information across a text to provide explanations, interpret significance and sequence activities

Source: PIRLS 2016 Reading Framework

Appendix B Data preparation and recoding

The PIRLS 2016 data retrieved from the IEA international database contains comprehensive student response data for each item in the PIRLS 2016 reading assessment. The response data details whether students answered correctly or incorrectly, which multiple choice option was selected and, for polytomously scored items whether their responses were scored as partially correct, or almost correct. When retrieved using EdSurvey, the responses in the dataset are presented in word/character form (e.g., CORRECT RESPONSE, ALMOST CORRECT RESPONSE) and need to be recoded to numerical responses (e.g., 0, 1, 2) to be analysed by the mirt package in R. Table A-5 describes the original response values for the items as well as the recoded response values.

Table A-5

Item response original response and recoded values

		Recoded value		
Original response value	**Description**	**Dichotomous items**	**Polytomous items**	
			Max 2	**Max 3**
INCORRECT RESPONSE	Student attempted the item and gave an incorrect response	0	0	0
PARTIALLY CORRECT RESPONSE	Student attempted the item and gave a partially correct response	-	1	1
ALMOST CORRECT RESPONSE	Student attempted the item and gave an almost correct response	-	-	2
CORRECT RESPONSE	Student attempted the item and gave a correct response	1	2	3
A, B, C, D	Student attempted a multiple-choice item and selected an incorrect option	0	-	-
A*, B*, C*, D*	Student attempted a multiple-choice item and selected the correct option	1	-	-
NOT REACHED	Student attempted previous items for this passage but did not reach the item (or any of the subsequent items)	[blank]		

Original response value	Description	Recoded value		
		Dichotomous items	Polytomous items	
			Max 2	Max 3
OMITTED OR INVALID	Student should have answered the item but did not, or student attempted to answer an item, but the answer was invalid (e.g., selected 2 multiple-choice options)	0		
NA	The item was not administered or not available to the student.*	[blank]		

*Note: * Possible reasons include: students were not presented with that passage by design, item was left out of during printing, item was removed for contextual reasons, or item was not administered to a particular subgroup/country as it was not internationally comparable*

In addition to the responses that were given by students for each item, the data also contains information about responses that were not given (i.e., missing). The description column for different types of missing data in Table A-5 shows the details provided in the PIRLS guide for missing responses. It is important to consider the treatment of missing data carefully in any analysis. This is particularly true for when using IRT models as the way in which missing responses are treated can influence both the item and person estimates (De Ayala et al., 2001; Finch, 2008; Lord, 1980). There are many different approaches to handling missing data, common options include considering missing responses as incorrect responses, as partially/fractionally correct responses, leaving the values as missing, or employing a combination.

The choice of which approach to apply should be made with care and account for the potential reason a student did not respond to an item. Often these reasons are unknown; for example, it is impossible to know exactly why a student made the choice to omit to respond to an item during an assessment. There are, however, cases where the reason may be known; for example, if a particular item was not administered to a certain group of students. In the latter case, often termed ‘structurally missing’, it would be inappropriate to consider the missing responses as incorrect as the reason the student did not respond is unrelated to their underlying ability on the construct. When the reason is unknown (i.e., if we are not sure whether the student missed the item, did not reach the item, or chose not to respond because they did not know the answer) assuming that it is not related to the construct may result in inflated item discrimination and item difficulty estimates, as well as inaccurate person ability estimates (Dai et al., 2021; Finch, 2008; Lord, 1980).

In the PIRLS Literacy 2016 response data, separate codes are provided for missing data that is structurally missing and data that was omitted by the student. Structurally missing data is coded as NOT ADMINISTERED (NA) if students were not presented with an item by design (i.e., that passage was not in their booklet) or if the item was left out of during printing; removed for contextual reasons; or not administered to a particular subgroup/country as it was not internationally comparable (Foy, 2018a). Missing data that was omitted by students for unknown reasons is either coded as OMITTED or NOT REACHED. Items coded as NOT REACHED in the PIRLS 2016 data are items assumed not to be attempted due to lack of time, in other words the students attempted the first few items for a particular passage but did not attempt a string of two or more consecutive items at the end of the passage. In these cases, the last answer attempted by a student is identified and scored, the following item is then coded as OMITTED and all subsequent items are coded as NOT REACHED (Martin, Mullis, & Hooper, 2017).

To avoid overestimating the item parameters, and in line with the approach used by the TIMSS & PIRLS International Study Center, this study chose to consider items which were OMITTED OR INVALID as INCORRECT RESPONSES (i.e., considered to be omitted for reasons related to the construct) while items that were NOT ADMINISTERED and NOT REACHED were left as missing/blank (i.e., considered to be omitted for reasons that may be unrelated to the construct). Similar to the response data, the missing data are presented in word/character form when retrieved using EdSurvey and need to be recoded to numerical responses (e.g., 0, 1, or [blank]).

The recoding of missing data was applied to all response data in the dataset. Table A-5 provides the original and recoded response values for the missing data and Table A-6 provides details regarding the number of missing responses that were recoded for each language group. In total across all item types, 1424 items originally coded as NOT REACHED were recoded as missing/blank and 5157 items originally coded as OMITTED OR INVALID were recoded as 0 (i.e., incorrect). Items that were not administered or available to the student (i.e., structurally missing) were coded as NA (not administered) in the original dataset and recoded as missing/blank to be analysed by mirt.

Table A-6

Occurrences of recoded missing data by language group

Missing data type	Afrikaans	English	Sepedi	Total
Not reached	329	744	351	1424
Omitted or invalid	1664	2116	1377	5157

After recoding the missing data, items that had null response categories needed to be recoded. Two polytomous items were found to have no responses in the CORRECT RESPONSE category. Item183 was originally a polytomous item with three categories (i.e., 0, 1, 2), however, no students were awarded full credit and so the item was treated as a dichotomous item (i.e., 0, 1). Item188 was originally a polytomous item with four categories (i.e., 0, 1, 2, 3), however, no students were awarded full credit and so the item was treated as a polytomous item with just three categories (i.e., 0, 1, 2). In both cases, no response values were changed, instead the assigned item-type for these two items was recoded to ensure that the items would be analysed using the appropriate IRT model. Once the required item response data was recoded, the dataset was ready to be analysed.

Appendix C Paper 1 additional information

This section presents some additional information relevant to Paper 1. Further details on the scaled items used from the contextual questionnaires are provided as well as a discussion of the assumption checks and diagnostics that were undertaken prior to conducting the multiple linear regression analysis.

C-1 Scaled items

Scaled items are values provided by PIRLS that represent sets of items that measure one underlying construct. For example, the ‘Students like reading’ derived scale asked students to respond to ten statements related to their attitude towards reading (e.g., “I would like to have more time for reading”, “I learn a lot from reading” and “I enjoy reading”). Students indicated their level of agreement with each statement on a four-point scale; “agree a lot,” “agree a little,” “disagree a little,” and “disagree a lot”. The responses to these item sets were combined into scales by PIRLS using the Rasch Partial Credit Model. These scaled values can be considered continuously or, to facilitate interpretation, they are divided categorically into high, middle, and low regions. For example, the students like reading scale is divided into “very much like reading”, “somewhat like reading” and “do not like reading” based on identified response combinations. Table A-7 provides a summary of the scaled items used in this analysis and the associated items, prompts and response options for each scale.

Table A-7

Summary of item sets for scaled items and scale reliability statistics

Variable	Prompt and item set	Response options	Cronbach’s alpha
Students sense of school belonging	**What do you think about your school? Tell how much you agree with these statements.** I like being in school I feel safe when I am at school I feel like I belong at this school	**Agree a lot; Agree a little; Disagree a little; Disagree a lot**	**0.74**

Variable	Prompt and item set	Response options	Cronbach's alpha
	Teachers at my school are fair to me I am proud to go to this school		
Student experience of bullying	**During this school year, how often have other students from your school done any of the following things to you (including through texting or the Internet)?** Made fun of me or called me names Left me out of their games or activities Spread lies about me Stole something from me Hit or hurt me Made me do things I didn't want to do Shared embarrassing information about me Threatened me	**At least once a week; Once or twice a month; A few times a year; Never**	**0.79**
Students engaged in reading lessons	**How much do you agree with these statements about your reading lessons?** I like what I read about in school My teacher gives me interesting things to read I know what my teacher expects me to do My teacher is easy to understand I am interested in what my teacher says My teacher encourages me to say what I think about what I have read My teacher lets me show what I have learned My teacher does a variety of things to help us learn My teacher tells me how to do better when I make a mistake	**Agree a lot; Agree a little; Disagree a little; Disagree a lot**	**0.84**
Students like reading	**What do you think about reading? Tell how much you agree with each of these statements.** I like talking about what I read with other people I would be happy if someone gave me a book as a present I think reading is boring*	**Agree a lot; Agree a little; Disagree a little; Disagree a lot**	**0.75**

Variable	Prompt and item set	Response options	Cronbach's alpha
	I would like to have more time for reading		
	I enjoy reading		
	I learn a lot from reading		
	I like to read things that make me think		
	I like it when a book helps me imagine other worlds		
Students confident in reading	**How well do you read? Tell how much you agree with each of these statements**	**Agree a lot; Agree a little; Disagree a little; Disagree a lot**	**0.58**
	I usually do well in reading		
	Reading is easy for me		
	I have trouble reading stories with difficult words*		
	Reading is harder for me than for many of my classmates*		
	Reading is harder for me than any other subject*		
	I am just not good at reading*		
Home resources for learning	**(Varied prompts)**		**0.62**
	Number of books in the home (students)	None or very few (0-10 books); Enough to fill one shelf (11-25 books); Enough to fill one bookcase (26-100 books); Enough to fill two bookcases (101-200 books); Enough to fill three or more bookcases (more than 200)	
	Number of children's books in the home (parents)	0-10; 11-25; 26-50; 51-100; More than 100	
	Number of home study supports (students)	Neither own room nor internet connection; Either own room or internet connection; Both own room and internet connection	
	Highest level of education of either parent (parents)	University or higher; Post-secondary but not university; Upper secondary; Lower secondary; Some primary, lower secondary or no school	

Variable	Prompt and item set	Response options	Cronbach's alpha
	Highest level of occupation of either parent (parents)	Professional; Small business owner; Clerical; Skilled worker; General labourer; Never worked for pay	
Early literacy activities at home before school	**Before your child began primary school, how often did you or someone else in your home do the following activities with him or her?**	**Often; Sometimes; Never or almost never**	**0.77**
	Read books		
	Tell stories		
	Sing songs		
	Play with alphabet toys		
	Talk about things you had done		
	Talk about things you had read		
	Play word games		
	Write letters or words		
	Read aloud signs and labels		
Could do early literacy tasks before school	**How well could your child do the following when he/she began the first grade of primary school?**	**Very well; Moderately well; Not very well; Not at all**	**0.88**
	Recognise most of the letters of the alphabet		
	Read some words		
	Read sentences		
	Read a story		
	Write letters of the alphabet		
	Write some words		
Parents like reading	**Please indicate how much you agree with the following statements about reading**	**Agree a lot; Agree a little; Disagree a little; Disagree a lot**	**0.75**
	I read only if I have to*		
	I like talking about what I read with other people		
	I like to spend my spare time reading		
	I read only if I need information*		
	Reading is an important activity in my home		
	I would like to have more time for reading		
	I enjoy reading		
	Reading is one of my favourite hobbies		

Variable	Prompt and item set	Response options	Cronbach's alpha
	When you are at home, how often do you read for your own enjoyment?	Every day or almost every day; Once or twice a week; Once or twice a month; Never or almost never	
Reading resource shortages	**How much is your school's capacity to provide instruction affected by a shortage or inadequacy of the following?**	**Not at all; A little; Some; A lot**	**0.81**
	Instructional materials		
	Supplies		
	School buildings and grounds		
	Heating/cooling and lighting systems		
	Instructional space		
	Technologically competent staff		
	Audio-visual resources for delivery of instruction		
	Computer technology for teaching and learning		
	Teachers with a specialisation in reading		
	Computer software/applications for reading instruction		
	Library resources		
	Instructional materials for reading		
School emphasis on academic success	**How would you characterise each of the following within your school?**	**Very high; High; Medium; Low; Very low**	**0.9**
	Teachers' understanding of the schools curricular goals		
	Teachers' degree of success in implementing the school's curriculum		
	Teachers' expectations for student achievement		
	Teachers' ability to inspire students		
	Collaboration between school leadership and teachers to plan instruction		
	Parental involvement in school activities		
	Parental commitment to ensure that students are ready to learn		
	Parental expectations for student achievement		
	Parental support for student achievement		

Variable	Prompt and item set	Response options	Cronbach's alpha
	Students' desire to do well in school Students' ability to reach school's academic goals Students' respect for classmates who excel academically		
School discipline	**To what degree is each of the following a problem among fourth grade students in your school?** Arriving late at school Absenteeism Classroom disturbance Cheating Profanity Vandalism Theft Intimidation or verbal abuse among students Physical fights among students Intimidation or verbal abuse of teachers or staff	**Not a problem; Minor problem; Moderate problem; Serious problem**	**0.9**

*Note: * indicates items that were reverse coded*

C-2 Assumptions and diagnostics of the regression model

The reliable estimation of multiple linear regression models and accurate interpretation of results depends on a number of assumptions about the data and model being met. Therefore, before proceeding with the analysis in this paper, diagnostic checks were conducted on the regression model to evaluate the assumptions of linearity, no multicollinearity, homoscedasticity, and normality of residuals (Osborne & Waters, 2019).

Linearity

A linear regression analysis relies on the assumption that there is a linear relationship between the dependent variable and each of the predictor variables when all other predictors are held constant (Cohen et al., 2013). In other words, in this paper the analysis assumes that there is a linear relationship between each of the contextual and language-related variables being used as predictor variables and reading achievement in PIRLS. This assumption is based on the relationship between reading achievement and these factors outlined by theoretical and empirical contributions of other research outlined in Section 5.2 in Paper 1. However, it is still necessary to inspect the relationship for each of the predictor variables individually before making inferences from the results of the regression analysis.

To assess linearity, the individual relationships between each predictor and the dependent variable are examined using partial-residual plots. These partial-residual plots represent the relationship between the predictor variable plus the residuals of the dependent variable (i.e., reading achievement scores). In each plot, the dashed line represents a regression line for the comparison pair and the solid line is a smoothed line where the linearity assumption is marginally relaxed. If the solid and dashed line fall together then the assumption holds well for that predictor. For categorical variables, the assumption of linearity is automatically satisfied as each category is compared with one reference category, the line therefore runs between two points and will inevitably be a straight line. The partial-residual plots, shown in Figure A-4, are therefore only examined for continuous variables. For two of these predictors (*acbgeas* and *asbhplr*) the linearity assumption holds. However, for *asbghrl* and *acbgrrs*, there is evidence that the relationship is not perfectly linear. These variables were therefore considered carefully to evaluate evidence of multicollinearity in subsequent diagnostic checks.

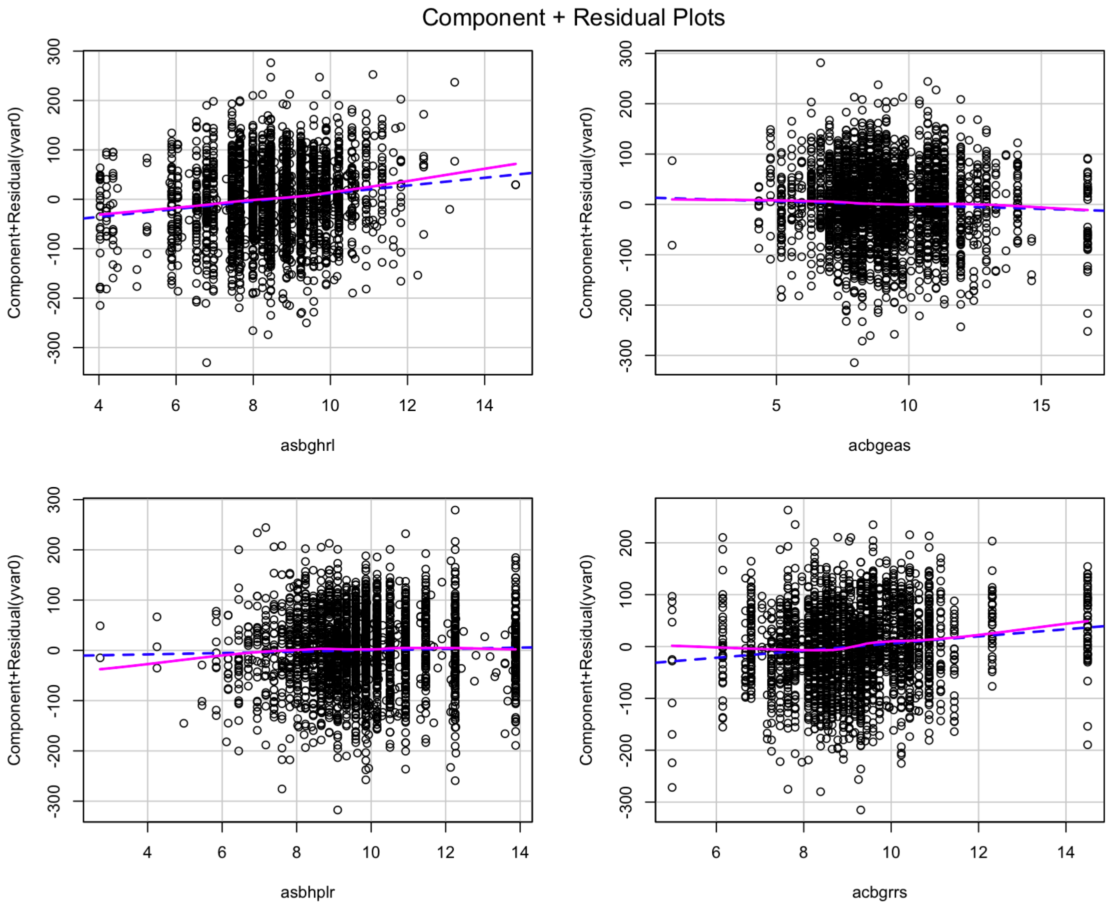


Figure A-4 Partial-residual plots for continuous predictor variables

A further inspection of linearity involves examining the relationship between the residuals from the regression model versus the fitted values for the model as a whole. In this plot, shown in Figure A-5, no discernible pattern should be evident in the data. More specifically, evidence of the assumption of linearity being upheld by the model occurs when the reference line is roughly horizontal at zero across all values (Osborne & Waters, 2019). As seen in the plot, there is no clear relationship between the residuals and the fitted values for the regression model. This means that the assumption of linearity holds for this model.

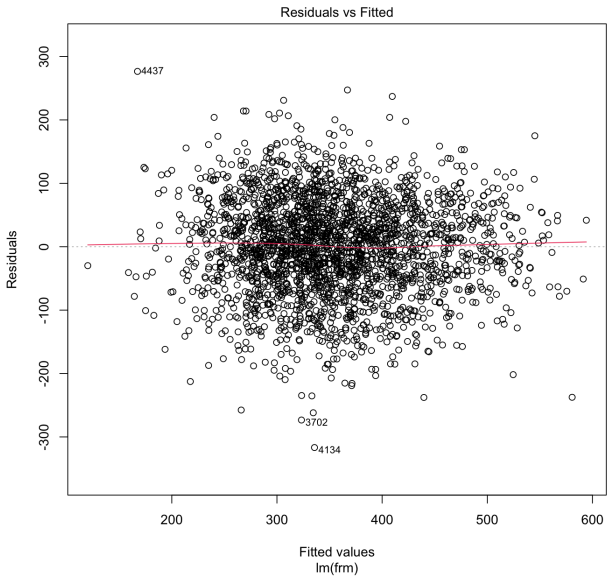


Figure A-5 Residuals of regression model versus fitted values for Model 5

Multicollinearity

One of the primary threats with a regression model with multiple predictor variables such as the one being used in this study is the threat of multicollinearity. Multicollinearity occurs when predictor variables are highly correlated with one another and can result in a number of negative consequences for the model, as well as interpreting the results. It can inflate standard errors; coefficient estimates as well as obscuring the influence of certain variables. These effects are mitigated to an extent through the use of the set of PVs as a dependent variable, as well as the JRR method recommended for survey data such as this (Bailey & Cohen, 2019; Foy, 2018a). However, it is still necessary to evaluate the relationships between predictor variables carefully in order to ensure the results are precise, accurate and reliable reflection of the relationship between background variables and reading achievement. To do this, the correlations between each pair of predictor variables are examined.

Figure A-6 presents the correlation matrix for the 26 predictor variables. Visual inspection of the correlations revealed no concerning issues , as all variable pairs exhibited correlation coefficients below the common threshold of .80 used to indicate high intercorrelation (Thompson et al., 2017). The strongest correlation was seen between the language that the mother (*asbh21ab*) and father (*asbh21aa*) use to communicate with their child at home at r = .60, which are variables which would be expected to have a relationship. Though these predictors share a substantial amount of variance, the association does not reach levels that necessitate removal due to redundancy. With all correlations in the reasonable range, all predictor variables were retained at this stage of analysis.

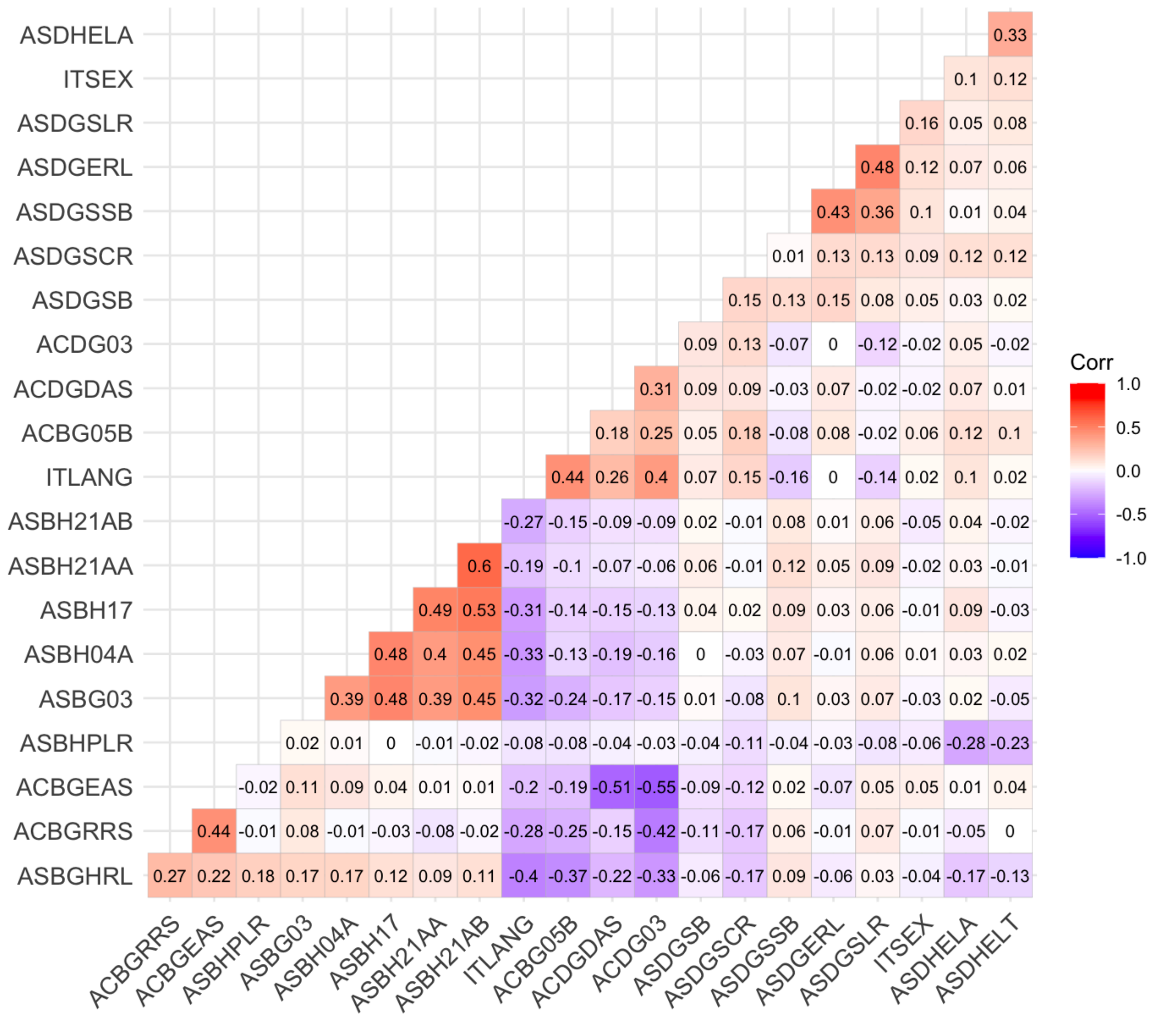


Figure A-6 Correlation matrix for predictor variables in Model 5

Although the relationship between pairs of predictor variables can be detected with a correlation matrix, multicollinearity is a function of the interrelationships between predictor variables as they occur simultaneously in the model, and therefore requires further examination. To evaluate multicollinearity, variance inflation factors (VIF) were calculated. VIFs for each predictor account for its correlation with all other predictors. In this way, VIF values can be used to quantify how much the variance of a regression coefficient is inflated due to multicollinearity in the model. VIF values above 10 are considered evidence of problematic collinearity between predictors (Thompson et al., 2017). Because this model includes multiple predictor variables, a generalised VIF value is also estimated that accounts for the number of parameters estimated (O'Brien, 2007). The results VIF values and generalised VIF values for the predictor variables are shown in Table A-8.

Table A-8

Variance inflation factors for variables in Model 5

Variable	GVIF	Df	GVIF^(1/(2*Df))
itsex	1.09	1	1.04
asbghrl	1.43	1	1.20
asdhela	1.32	2	1.07
asdhelt	1.28	2	1.06
asbhplr	1.17	1	1.08
asdgssb	1.53	2	1.11
asdgsb	1.16	2	1.04
asdgerl	1.67	2	1.14
asdgslr	1.59	2	1.12
asdgscr	1.24	2	1.06
acbgrrs	1.59	1	1.26
acbgeas	2.40	1	1.55
acdgdas	1.98	2	1.19
acbg05b	3.48	4	1.17
acdg03	1.62	1	1.27
asbg03	1.53	2	1.11
asbh17	1.64	2	1.13
asbh21aa	1.73	1	1.32
asbh21ab	1.92	1	1.39
asbh04a	1.57	1	1.25
itlang	5.51	10	1.09

The highest individual VIF value is 5.51 for itlang, while all other predictors have values below 5. Furthermore, the square root of the VIF divided by the degrees of freedom for each variable provides an index that accounts for the number of parameters estimated. The highest value on this index is 1.55 for acbgeas. These VIF results confirm the lack of serious multicollinearity initially suggested through examination of the correlation matrix.

Homoscedasticity

Homoscedasticity assumes that the residuals should show constant variance across the range of all predicted values. In order to assess the homoscedasticity in a regression model, the predicted values are plotted against the standardised residual values in a scale-location plot. The scale-location plot is then examined to determine if the points are equally distributed across all values of the predictor variables. Figure A-7 shows the scale-location plot for the model. If homoscedasticity was present there would be a random distribution of points along the full range of fitted values and a flat horizontal line. However, as evident in the plot for this model, there is a slight curve of the line, indicating that there may be heteroscedasticity in this data. This means that further evaluation of the potential heteroscedasticity is necessary. This was undertaken using Penn and Slate (2006) approach which tests whether the residual variance is relatively constant across a range of values. The test for heteroscedasticity showed that there was no significant difference across the residual variances at the $p > .05$ level ($p = .739$). In other words, the assumption of homoscedasticity is acceptable for this model.

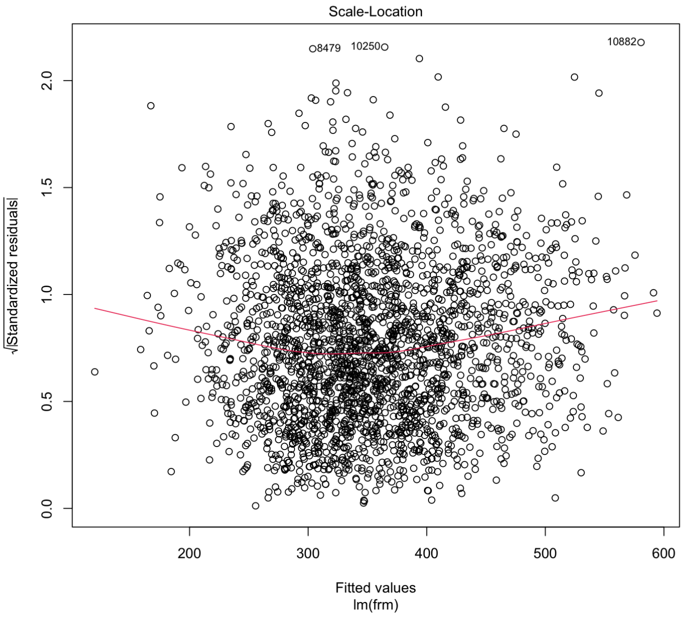


Figure A-7 Scale-location plot of standardised residuals for Model 5

Normality

The assumption of normality requires residuals to be normally distributed around a mean of zero. To assess normality, a histogram and Q-Q plot of the residuals were reviewed. The histogram in Figure A-8 shows evidence of a normal distribution of the residuals.

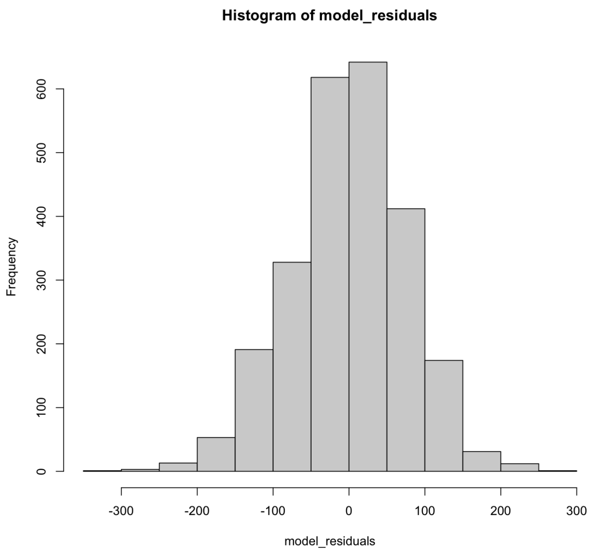


Figure A-8 Histogram of model residuals for Model 5

Next, we consider the residuals along a normal line using a QQ-plot. In the normal QQ-plot shown in Figure A-9, the majority of the residuals fall along the normal line showing further evidence of a normal distribution of residuals for this model.

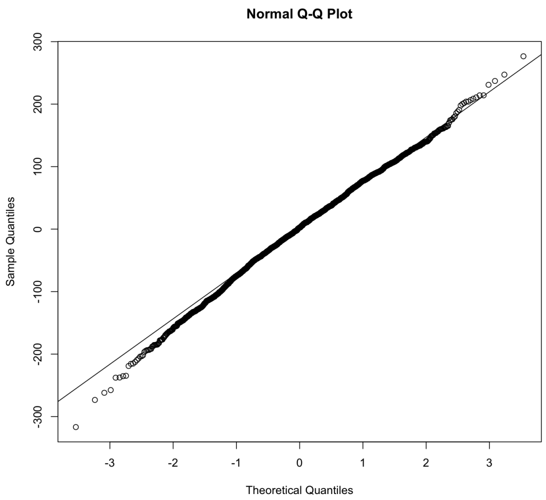


Figure A-9 Normal QQ-plot of model residuals for Model 5

Summary of diagnostic checks

The diagnostic checks conducted therefore indicate that the key assumptions of linearity, lack of multicollinearity, homoscedasticity, and normality of residuals were sufficiently met to justify proceeding with the multiple linear regression analysis in Paper 1. While not all assumptions were satisfied perfectly, no serious violations were detected that would drastically undermine the reliability of the regression results. Linearity was confirmed through examination of partial residual plots and residuals vs fitted values. The correlation matrix and variance inflation factors provided assurance that multicollinearity was not a major issue among the predictor variables. Some potential heteroscedasticity was initially flagged in the scale-location plot, but a formal test indicated constant residual variance was acceptable. Normality of residuals could not be stated definitively, as the QQ-plot showed some departure from the normal line despite the histogram suggesting an overall normal distribution. This provides reasonable grounds for treating the model's estimates as reliable representations of how background contextual factors relate to reading performance.

C-3 Detailed result tables for regression Models 0 to 5

Table A-9

Results of multiple linear regression Model 0

Model 0 variables	B	β	t	
Intercept	346.71		84.92	***
Model 0 R^2 = 0.06				
CONTROL FACTOR				
Student gender				
(Girl)				
Boy	-52.11	-0.24	-17.03	***

Notes:
( ) reference category
B regression coefficients
β standardised regression coefficients
t t-value
*Significance codes p < .001 '***'; p < .01 '**' ; p < .05 '*'*

Table A-10

Results of multiple linear regression Model 1

Model 1 variables	B	β	t	
Intercept	126.51		4.33	***
Model 1 R^2 = 0.23				
CONTROL FACTOR				
Student gender				
(Girl)				
Boy	-43.33	-0.19	-10.31	***
HOME FACTORS				
Home resources for learning				
	26.69	0.36	9.3	***
Early literacy activities at home				
(Often)				
Sometimes	-3.88	-0.02	-0.73	
Never or almost never	-27.04	-0.05	-2.33	*
Could do early literacy tasks before school				
(Very well)				
Moderately well	-9.2	-0.04	-1.82	
Not well	-37.82	-0.15	-5.33	***
Parents like reading				
	3.48	0.05	2.59	*

Notes:
( ) reference category
B regression coefficients
β standardised regression coefficients
t t-value
*Significance codes p < .001 '***'; p < .01 '**' ; p < .05 '*'*

Table A-11

Results of multiple linear regression Model 2

Model 2 variables	B	β	t	
Intercept	243.17		8.74	***
Model 2 R^2 = 0.35				
CONTROL FACTOR				
Student gender				
(Girl)				
Boy	-32.96	-0.15	-8.57	***
HOME FACTORS				
Home resources for learning				
	22.06	0.3	8.72	***
Early literacy activities at home				
(Often)				
Sometimes	-4.99	-0.02	-1.08	
Never or almost never	-25.44	-0.04	-2.23	*
Could do early literacy tasks before school				
(Very well)				
Moderately well	-4.03	-0.02	-0.83	
Not well	-27.55	-0.11	-4.35	***
Parents like reading				
	2.55	0.04	1.97	
STUDENT FACTORS				
Students sense of school belonging				
(High sense of school belonging)				
Some sense of school belonging	4.75	0.02	1.01	
Little sense of school belonging	7.17	0.02	0.83	
Student experience of bullying				

Model 2 variables	B	β	t	
(Almost never)				
About monthly	-11.99	-0.05	-2.24	*
About weekly	-31.37	-0.14	-7.42	***
Student engaged in reading lessons				
(Very engaged)				
Somewhat engaged	-20.14	-0.08	-3.27	**
Less than engaged	-44.14	-0.09	-4.55	***
Student likes reading				
(Very much like reading)				
Somewhat like reading	1.94	0.01	0.35	
Do not like reading	-0.44	0	-0.05	
Student confident in reading				
(Very confident)				
Somewhat confident	-47.76	-0.2	-7.15	***
Not confident	-83.53	-0.37	-10.58	***

Notes:
( ) reference category
B regression coefficients
β standardised regression coefficients
t t-value
*Significance codes p < .001 '***'; p < .01 '**' ; p < .05 '*'*

Table A-12

Results of multiple linear regression Model 3

Model 3 variables	B	β	t	
Intercept	295.34		6.39	***
Model 3 R^2 = 0.41				
CONTROL FACTOR				
Student gender				
(Girl)				
Boy	-31.51	-0.14	-8.32	***
HOME FACTORS				
Home resources for learning				
	10.36	0.14	5.11	***
Early literacy activities at home				
(Often)				
Sometimes	-5.04	-0.02	-1.19	
Never or almost never	-20.74	-0.04	-1.84	
Could do early literacy tasks before school				
(Very well)				
Moderately well	-9.35	-0.04	-2.29	*
Not well	-33.81	-0.14	-6.54	***
Parents like reading				
	2.63	0.04	1.99	
STUDENT FACTORS				
Students sense of school belonging				
(High sense of school belonging)				
Some sense of school belonging	-0.01	0	0	
Little sense of school belonging	-0.86	0	-0.09	
Student experience of bullying				

Model 3 variables	B	β	t	
(Almost never)				
About monthly	-2.71	-0.01	-0.59	
About weekly	-27.45	-0.12	-6.53	***
Student engaged in reading lessons				
(Very engaged)				
Somewhat engaged	-12.72	-0.05	-1.93	
Less than engaged	-32.82	-0.07	-3.13	**
Student likes reading				
(Very much like reading)				
Somewhat like reading	-5.17	-0.02	-0.89	
Do not like reading	-12.64	-0.03	-1.45	
Student confident in reading				
(Very confident)				
Somewhat confident	-34.09	-0.15	-4.72	***
Not confident	-61.63	-0.28	-8.03	***
SCHOOL FACTORS				
Reading resource shortages				
	10.76	0.15	3.65	***
School emphasis on academic success				
	-1.11	-0.02	-0.54	
School discipline				
(Hardly any problems)				
Minor problems	-5	-0.02	-0.38	
Moderate to severe problems	-18.07	-0.07	-1.36	
School location				
(Urban-densely populated)				
Medium size city or large town	20.23	0.04	0.72	

Model 3 variables	B	β	t	
Small town or village	-52.66	-0.19	-3.21	**
Remote rural	-64.55	-0.28	-4.03	***
Suburban/township	-27.12	-0.11	-1.47	
School composition by student background				
(Not disadvantaged)				
Disadvantaged	-16.92	-0.07	-1.6	

Notes:
( ) reference category
B regression coefficients
β standardised regression coefficients
t t-value
*Significance codes p < .001 '***'; p < .01 '**' ; p < .05 '*'*

Table A-13

Results of multiple linear regression Model 4

Model 4 variables	B	β	t	
Intercept	302.16		5.73	***
Model 4 R^2 = 0.45				
CONTROL FACTOR				
Student gender				
(Girl)				
Boy	-29.3	-0.13	-5.91	***
HOME FACTORS				
Home resources for learning				
	11.66	0.15	4.38	***
Early literacy activities at home				
(Often)				
Sometimes	-11.62	-0.05	-2.33	*
Never or almost never	-23.13	-0.04	-1.66	
Could do early literacy tasks before school				
(Very well)				
Moderately well	-8.14	-0.04	-1.52	
Not well	-31.49	-0.12	-5.21	***
Parents like reading				
	2.58	0.04	1.44	
STUDENT FACTORS				
Students sense of school belonging				
(High sense of school belonging)				
Some sense of school belonging	2.91	0.01	0.5	
Little sense of school belonging	3.11	0.01	0.22	
Student experience of bullying				

Model 4 variables	B	β	t	
(Almost never)				
About monthly	-8.27	-0.04	-1.28	
About weekly	-31.97	-0.14	-5.55	***
Student engaged in reading lessons				
(Very engaged)				
Somewhat engaged	-7.64	-0.03	-1.07	
Less than engaged	-32.53	-0.06	-2.07	*
Student likes reading				
(Very much like reading)				
Somewhat like reading	-1.23	-0.01	-0.16	
Do not like reading	-13.12	-0.03	-1.12	
Student confident in reading				
(Very confident)				
Somewhat confident	-32.77	-0.14	-4.44	***
Not confident	-60.96	-0.27	-8.02	***
SCHOOL FACTORS				
Reading resource shortages				
	10.26	0.15	3.05	**
School emphasis on academic success				
	-1.83	-0.04	-0.82	
School discipline				
(Hardly any problems)				
Minor problems	7.27	0.03	0.53	
Moderate to severe problems	-8.49	-0.03	-0.62	
School location				
(Urban-densely populated)				
Medium size city or large town	10.83	0.02	0.37	

Model 4 variables	B	β	t	
Small town or village	-58.96	-0.2	-3.3	**
Remote rural	-69.37	-0.28	-4.23	***
Suburban/township	-36.13	-0.15	-1.97	
School composition by student background				
(Not disadvantaged)	-19.36	-0.08	-1.58	
Disadvantaged	-1.83	-0.04	-0.82	
LANGUAGE USE FACTORS				
Test language spoken at home (student)				
(Always/almost always)				
Sometimes	13.77	0.05	1.56	
Never	-33.58	-0.07	-2.79	**
Child speaks test language at home				
(Always/almost always)				
Sometimes	11.36	0.04	1.44	
Never	27.38	0.02	1.69	
Father speaks test language with child				
(Yes)				
No	-14.31	-0.05	-2.38	*
Mother speaks test language with child				
(Yes)				
No	-5.77	-0.02	-0.7	
Child spoke test language before school				
(Yes)				
No	17.73	0.06	2.24	*

Notes:
( ) reference category
B regression coefficients
β standardised regression coefficients
t t-value
*Significance codes p < .001 '***'; p < .01 '**' ; p < .05 '*'*

Table A-14

Results of multiple linear regression Model 5

Model 5 variables	B	β	t	
Intercept	375.21		10.6	***
Model R^2 = 0.54				
CONTROL FACTOR				
Student gender				
(Girl)				
Boy	-27.08	-0.12	-6.2	***
HOME FACTORS				
Home resources for learning				
	8.48	0.11	4.29	***
Early literacy activities at home				
(Often)				
Sometimes	-5.9	-0.03	-1.26	
Never or almost never	-9.29	-0.02	-0.63	
Could do early literacy tasks before school				
(Very well)				
Moderately well	-10.97	-0.05	-2.11	*
Not well	-33.25	-0.13	-6.22	***
Parents like reading				
	1.75	0.03	1.15	
STUDENT FACTORS				
Students sense of school belonging				
(High sense of school belonging)				
Some sense of school belonging	-2.49	-0.01	-0.55	
Little sense of school belonging	-9.39	-0.02	-0.72	
Student experience of bullying				

Model 5 variables	B	β	t	
(Almost never)				
About monthly	-5.18	-0.02	-0.88	
About weekly	-27.97	-0.12	-4.37	***
Student engaged in reading lessons				
(Very engaged)				
Somewhat engaged	-10.85	-0.04	-1.62	
Less than engaged	-28.27	-0.05	-2.02	*
Student likes reading				
(Very much like reading)				
Somewhat like reading	-7.51	-0.03	-1.21	
Do not like reading	-22.92	-0.06	-2.08	*
Student confident in reading				
(Very confident)				
Somewhat confident	-29.36	-0.12	-4.32	***
Not confident	-60.53	-0.26	-9.61	***
SCHOOL FACTORS				
Reading resource shortages				
	8.88	0.13	2.82	**
School emphasis on academic success				
	-1.52	-0.03	-0.83	
School discipline				
(Hardly any problems)				
Minor problems	10.87	0.05	1.18	
Moderate to severe problems	0.09	0	0.01	
School location				
(Urban-densely populated)				
Medium size city or large town	4.32	0.01	0.18	

Model 5 variables	**B**	**β**	**t**	
Small town or village	-43.58	-0.15	-2.9	**
Remote rural	-34.58	-0.14	-2.51	*
Suburban/township	-23.29	-0.1	-1.59	
School composition by student background				
(Not disadvantaged)				
Disadvantaged	-14.52	-0.06	-1.71	
LANGUAGE USE FACTORS				
Test language spoken at home (student)				
(Always/almost always)				
Sometimes	5.79	0.02	0.95	
Never	-29.39	-0.06	-2.96	**
Child speaks test language at home				
(Always/almost always)				
Sometimes	1.63	0.01	0.19	
Never	3.61	0	0.22	
Father speaks test language with child				
(Yes)				
No	-11.87	-0.04	-2.15	*
Mother speaks test language with child				
(Yes)				
No	-13.23	-0.05	-1.72	
Child spoke test language before beginning school				
(Yes)				
No	3.36	0.01	0.44	
TEST LANGUAGE FACTOR				
Test language				

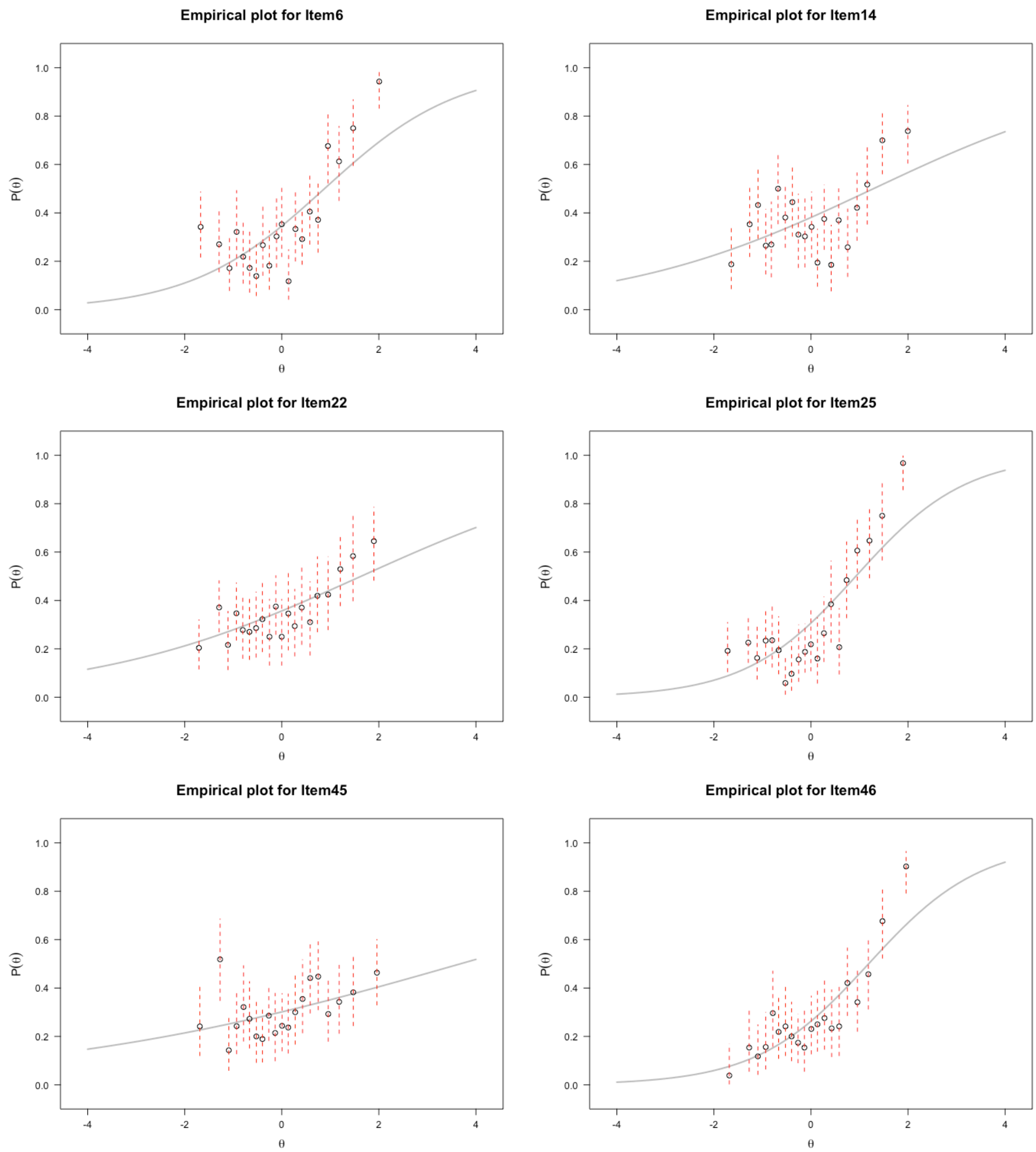
Model 5 variables	B	β	t	
(English)				
Afrikaans	-10.96	-0.03	-0.79	
isiNdebele	-64.36	-0.04	-4.88	***
isiXhosa	-66.48	-0.21	-4.5	***
isiZulu	-80.09	-0.25	-5.08	***
Sepedi	-130.25	-0.36	-10.98	***
Sesotho	-51.46	-0.1	-3.08	**
Setswana	-65.18	-0.13	-3.97	***
siSwati	-55.48	-0.07	-4.09	***
Tshivenda	-63.05	-0.08	-3.9	***
Xitsonga	-57.65	-0.07	-3.43	**

Notes:
( ) reference category
B regression coefficients
β standardised regression coefficients
t t-value
*Significance codes p < .001 '***'; p < .01 '**' ; p < .05 '*'*

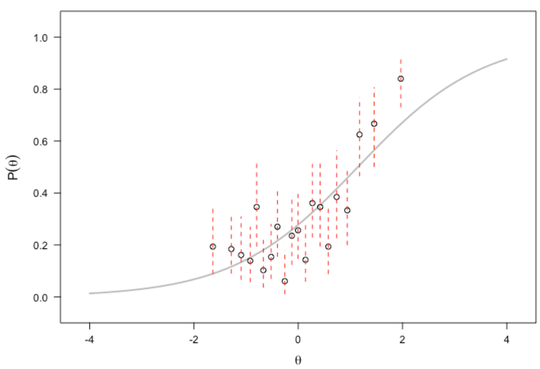
Appendix D Paper 2 additional information

Paper 2 presents selected examples of the IRFs for items that were shown to be misfitting or those items with evidence of negative discrimination parameters. This section presents figures for all items identified as misfitting due to graphical check of observed versus expected IRFs (Figure A-10) as well as those identified as having negative discrimination parameters in one or more groups (Figure A-11). Also included in this section are the item parameters for all three groups as freely estimated model (M_0), presented in Table A-15.

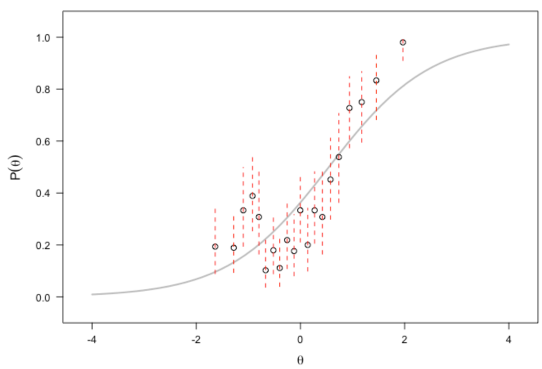
D-1 Item response functions for misfitting items



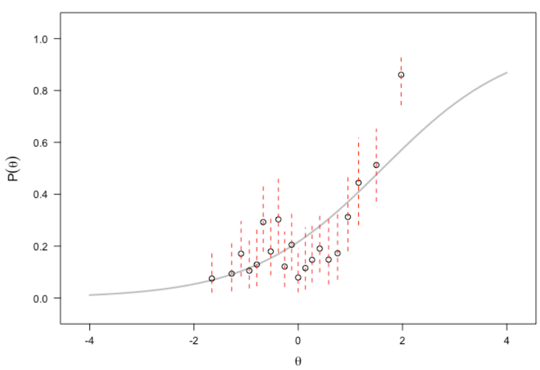
Empirical plot for Item49



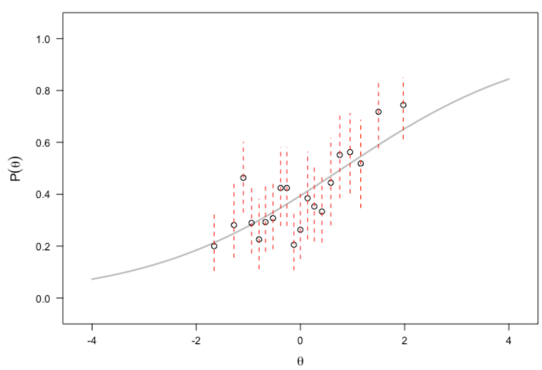
Empirical plot for Item53



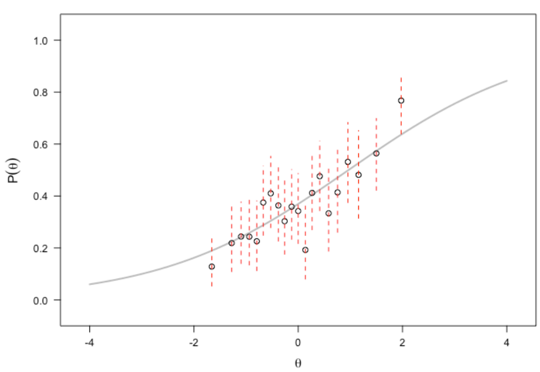
Empirical plot for Item65



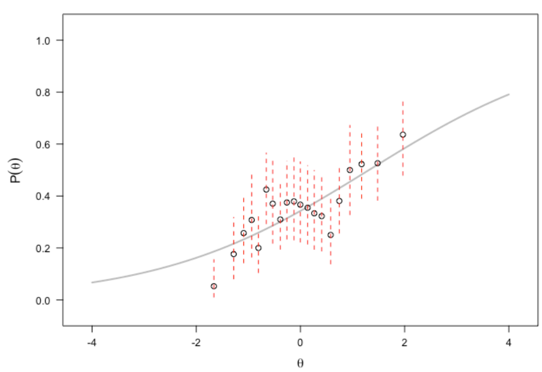
Empirical plot for Item66



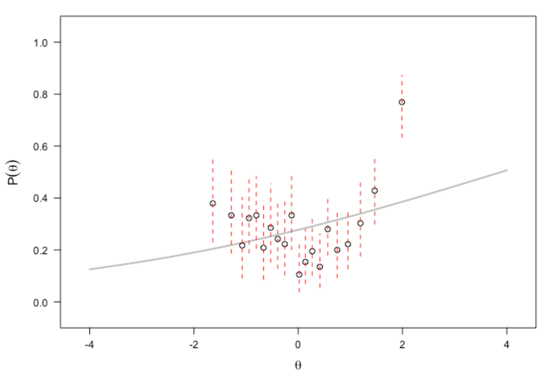
Empirical plot for Item70



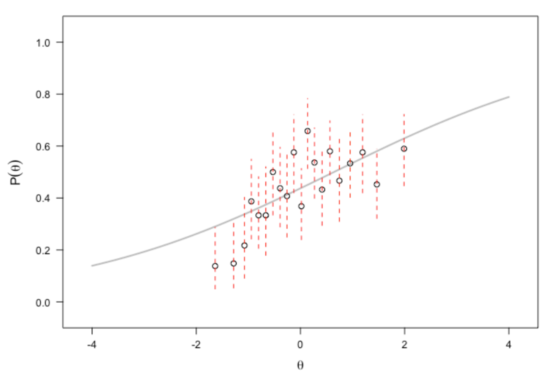
Empirical plot for Item79



Empirical plot for Item85



Empirical plot for Item86



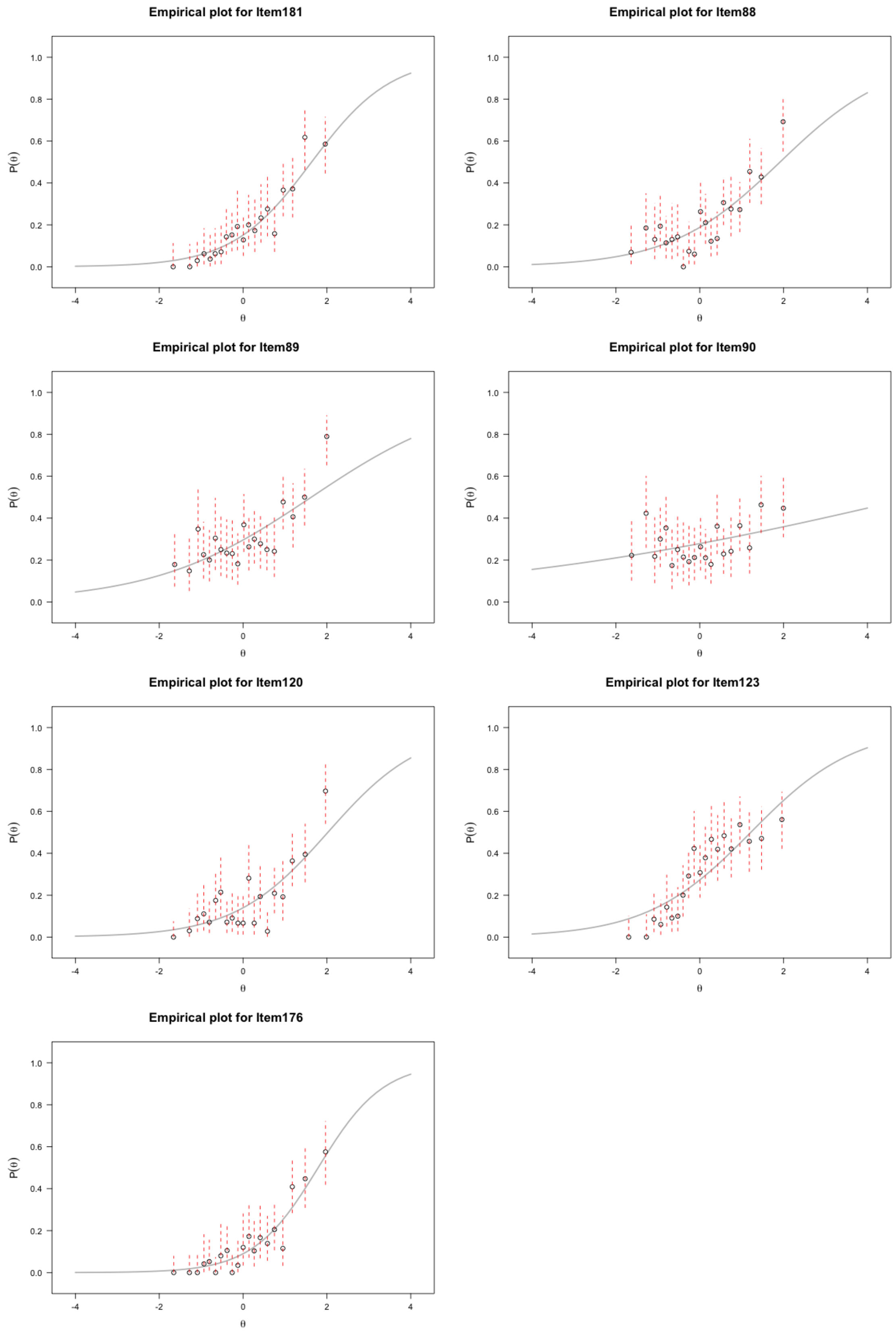
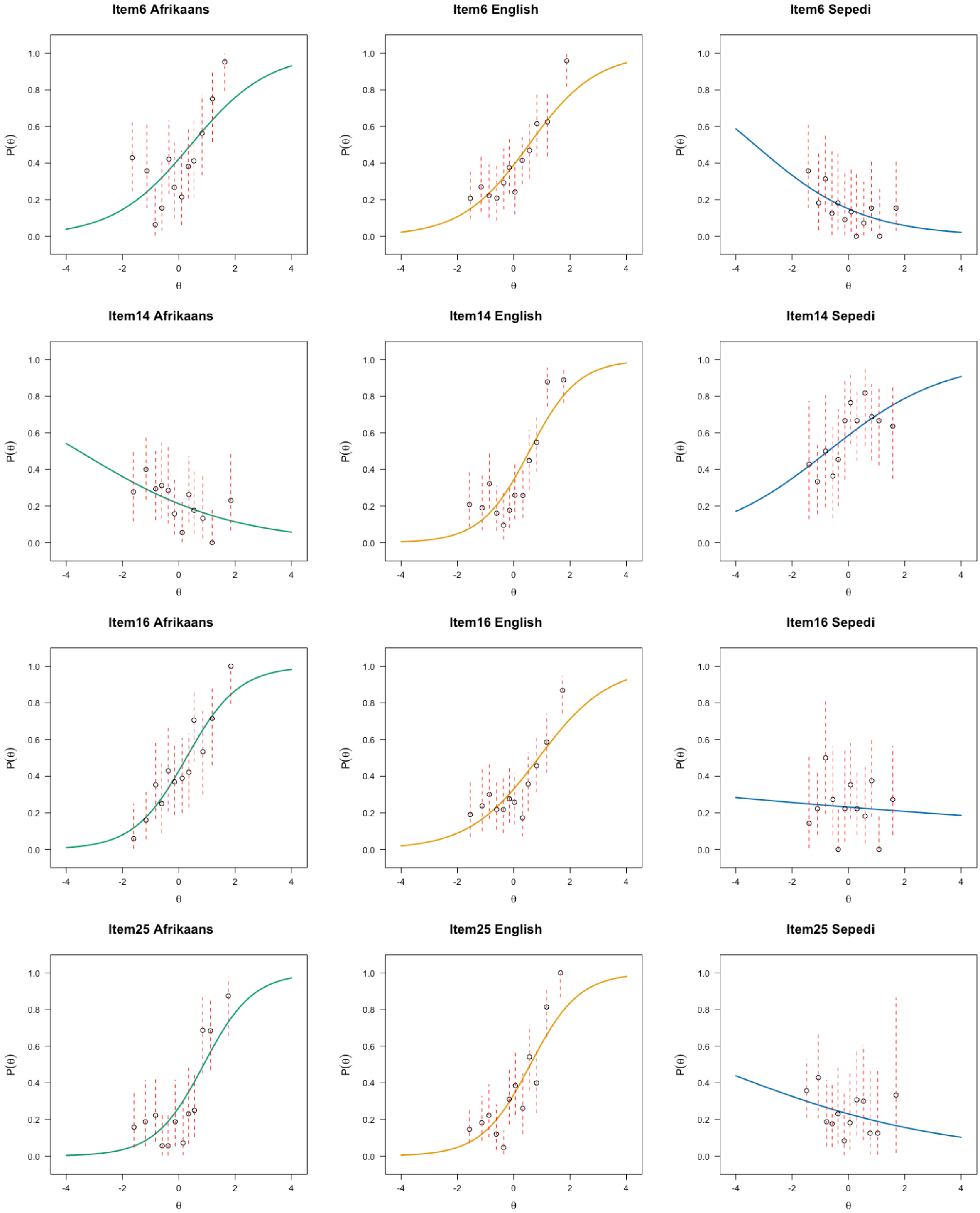
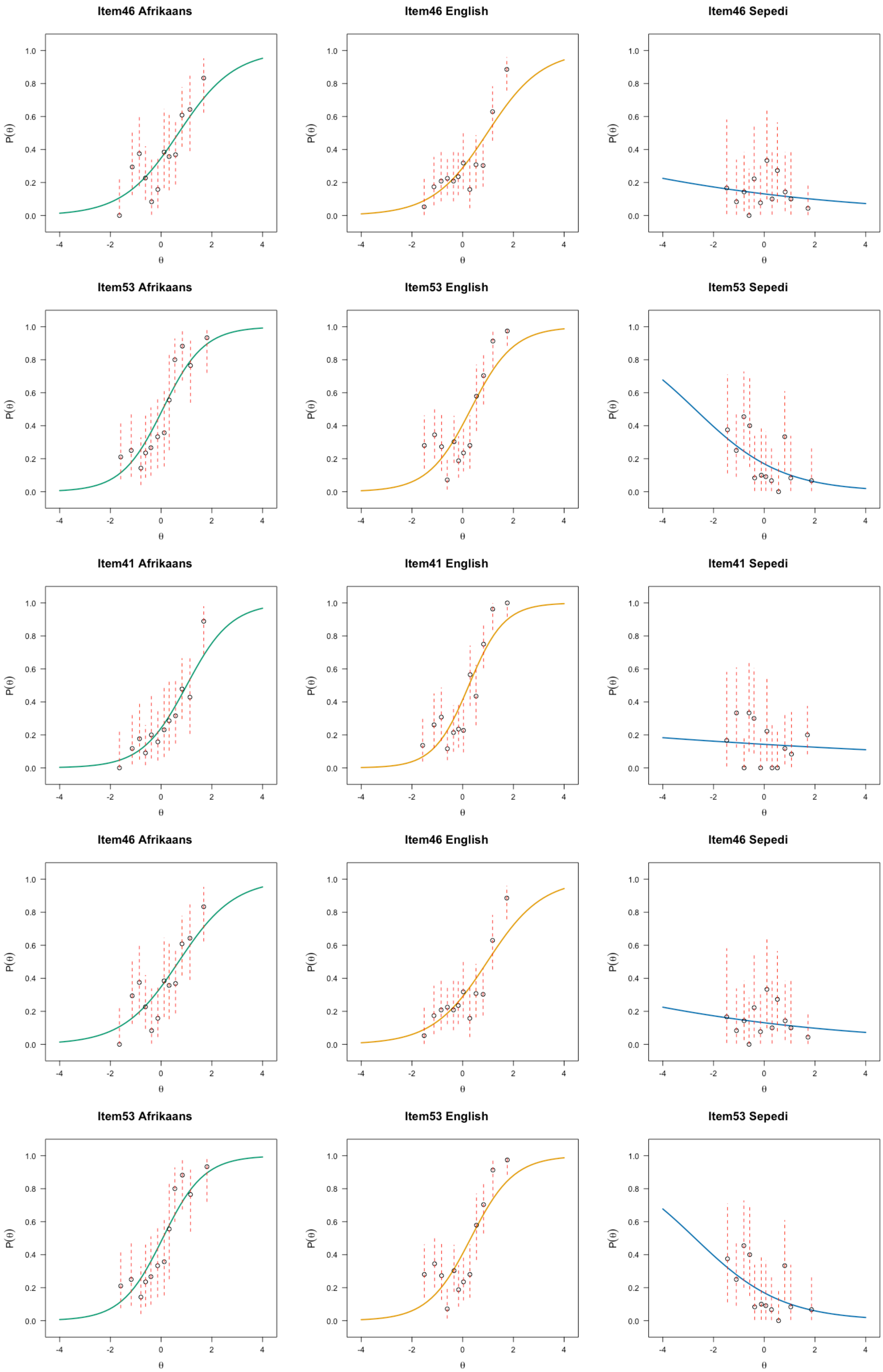


Figure A-10 Expected versus observed item response functions for 21 misfitting items

D-2 Item response functions for negatively discriminating items




Item46 Afrikaans
Item46 English
Item46 Sepedi
Item53 Afrikaans
Item53 English
Item53 Sepedi
Item41 Afrikaans
Item41 English
Item41 Sepedi
Item46 Afrikaans
Item46 English
Item46 Sepedi
Item53 Afrikaans
Item53 English
Item53 Sepedi
P(θ)
θ
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4

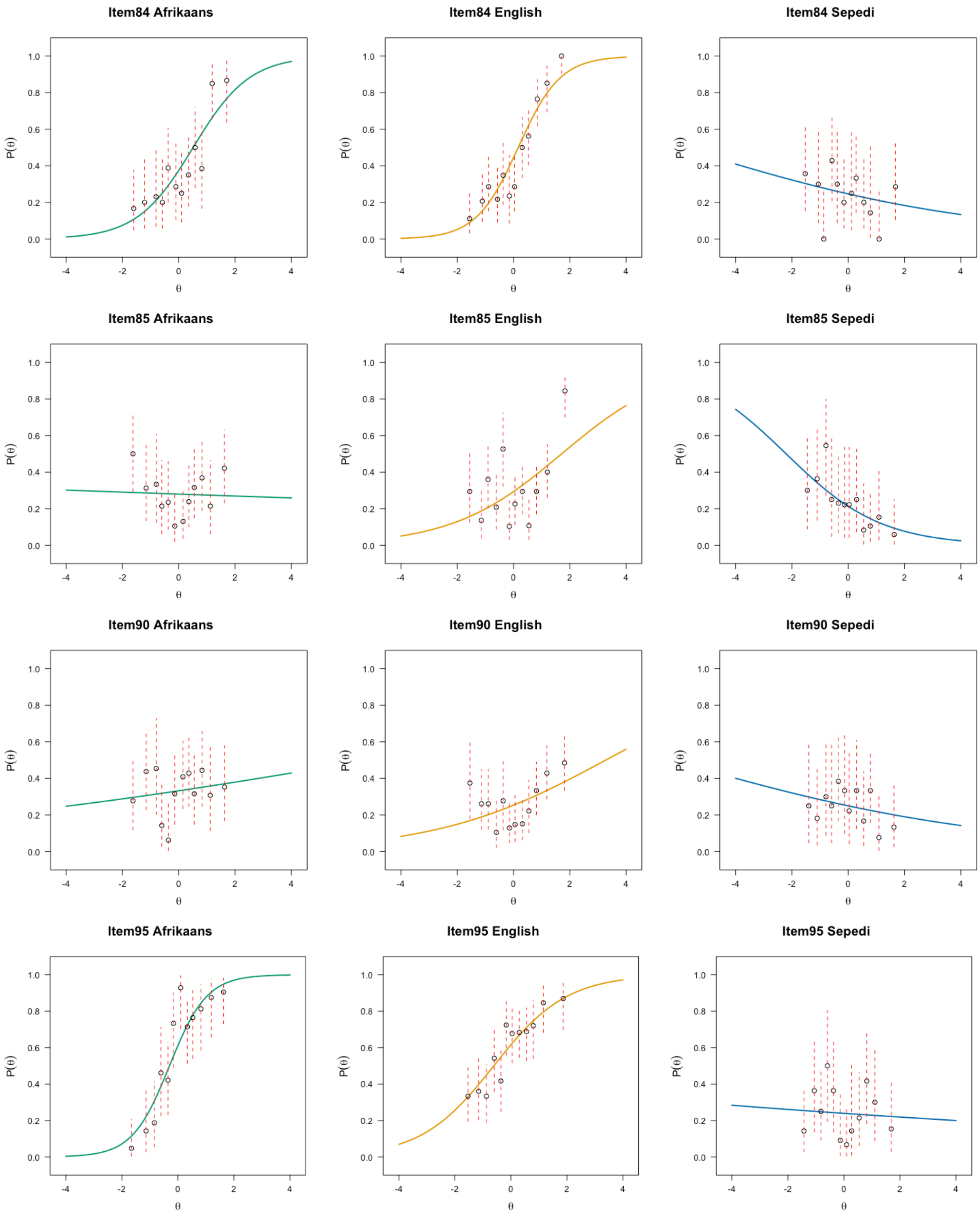


Figure A-11 Expected versus observed item response functions for negatively discriminating items across groups

Table A-15

Item parameters for freely estimated model

	English				Afrikaans				Sepedi			
	a	**d**	**d1**	**d2**	**a**	**d**	**d1**	**d2**	**a**	**d**	**d1**	**d2**
Item1	0.556	1.666	-	-	1.121	1.448	-	-	0.894	1.410	-	-
Item2	0.970	-0.618	-	-	0.530	-1.071	-	-	0.229	-1.362	-	-
Item3	2.490	2.245	-	-	1.579	1.812	-	-	1.164	0.626	-	-
Item4	1.999	0.422	-	-	1.657	0.074	-	-	0.971	-0.132	-	-
Item5	1.262	0.752	-	-	1.607	0.518	-	-	0.482	-0.800	-	-
Item6	0.836	-0.450	-	-	0.728	-0.312	-	-	-0.523	-1.744	-	-
Item7	2.505	2.077	-	-	3.284	1.618	-	-	1.454	0.355	-	-
Item8	1.067	0.228	-	-	0.913	0.266	-	-	0.429	-0.663	-	-
Item9	1.534	-0.314	-	-	1.932	-1.012	-	-	0.134	-0.858	-	-
Item10	2.101	2.171	-	-	2.043	1.815	-	-	-0.034	-0.343	-	-
Item11	3.777	4.293	-	-	1.711	2.055	-	-	1.322	1.136	-	-
Item12	2.305	1.503	-	-	1.878	0.113	-	-	1.363	0.483	-	-
Item13	1.257	0.581	-	-	0.718	0.258	-	-	0.257	0.033	-	-
Item14	1.163	-0.647	-	-	-0.370	-1.311	-	-	0.482	0.352	-	-
Item15	1.352	1.108	-	-	1.618	0.876	-	-	1.036	0.102	-	-
Item16	0.807	-0.711	-	-	1.079	-0.284	-	-	-0.069	-1.202	-	-
Item17	1.295	-0.219	-	-	1.332	-0.861	-	-	0.492	-1.538	-	-
Item18	2.566	2.495	-	-	2.981	2.096	-	-	2.049	1.017	-	-

	English				Afrikaans				Sepedi			
	a	**d**	**d1**	**d2**	**a**	**d**	**d1**	**d2**	**a**	**d**	**d1**	**d2**
Item19	1.365	1.521	-	-	2.884	2.514	-	-	1.318	-0.331	-	-
Item20	1.646	0.848	-	-	1.228	0.403	-	-	0.371	-0.495	-	-
Item21	1.109	0.667	-	-	1.061	0.399	-	-	0.333	0.257	-	-
Item22	0.437	-0.616	-	-	0.202	-0.287	-	-	0.086	-1.017	-	-
Item23	1.514	0.243	-	-	1.699	0.567	-	-	0.379	-0.226	-	-
Item24	1.507	1.553	-	-	1.443	0.707	-	-	1.252	0.649	-	-
Item25	1.159	-0.685	-	-	1.153	-1.017	-	-	-0.240	-1.208	-	-
Item26	2.392	2.625	-	-	2.860	1.881	-	-	1.593	0.767	-	-
Item27	1.696	0.693	-	-	1.656	0.132	-	-	0.931	-0.699	-	-
Item28	1.528	0.531	-	-	1.866	0.321	-	-	0.886	-0.620	-	-
Item29	1.946	0.765	-	-	1.604	0.471	-	-	0.597	-0.911	-	-
Item30	0.994	-0.510	-	-	0.864	-0.456	-	-	0.759	-1.287	-	-
Item31	1.531	-0.024	-	-	2.210	-0.921	-	-	-0.031	-0.829	-	-
Item32	1.124	0.502	-	-	1.611	0.068	-	-	0.440	-0.699	-	-
Item33	1.532	0.557	-	-	1.274	0.655	-	-	0.862	-0.508	-	-
Item34	1.856	1.174	-	-	1.810	0.984	-	-	0.685	-0.636	-	-
Item35	1.272	0.503	-	-	1.176	0.262	-	-	0.072	-1.047	-	-
Item36	1.635	0.814	-	-	1.683	0.157	-	-	0.650	-0.758	-	-
Item37	0.914	0.045	-	-	0.887	0.002	-	-	1.237	-0.124	-	-
Item38	1.390	0.383	-	-	1.169	0.080	-	-	0.867	-0.624	-	-
Item39	1.731	0.736	-	-	3.059	-0.327	-	-	0.670	-0.853	-	-

	English				Afrikaans				Sepedi			
	a	d	d1	d2	a	d	d1	d2	a	d	d1	d2
Item40	2.631	2.029	-	-	1.858	0.708	-	-	1.562	-0.970	-	-
Item41	1.444	-0.351	-	-	1.138	-1.145	-	-	-0.074	-1.793	-	-
Item42	1.147	0.397	-	-	1.178	0.160	-	-	1.233	-1.275	-	-
Item43	1.700	-0.191	-	-	1.197	-0.424	-	-	0.212	-1.573	-	-
Item44	1.767	0.082	-	-	1.708	-0.624	-	-	0.700	-0.895	-	-
Item45	0.020	-1.246	-	-	0.825	-0.399	-	-	0.239	-0.719	-	-
Item46	0.928	-0.898	-	-	0.910	-0.626	-	-	-0.164	-1.893	-	-
Item47	0.939	-0.371	-	-	0.837	-0.868	-	-	0.009	-1.292	-	-
Item48	1.989	1.592	-	-	2.123	0.981	-	-	0.575	-0.652	-	-
Item49	0.920	-0.841	-	-	1.188	-1.051	-	-	0.009	-1.100	-	-
Item50	1.660	1.666	-	-	1.302	1.003	-	-	0.392	-0.736	-	-
Item51	2.054	1.675	-	-	2.712	2.217	-	-	0.446	0.087	-	-
Item52	1.170	0.275	-	-	1.203	-0.634	-	-	0.789	-0.760	-	-
Item53	1.187	-0.379	-	-	1.232	-0.079	-	-	-0.584	-1.593	-	-
Item54	2.023	-0.291	-	-	2.475	-0.678	-	-	0.879	-1.181	-	-
Item55	1.063	0.480	-	-	1.331	-0.419	-	-	0.538	-1.167	-	-
Item56	1.822	0.406	-	-	1.331	0.161	-	-	0.029	-1.265	-	-
Item57	0.777	-0.201	-	-	0.541	-0.508	-	-	0.105	-1.003	-	-
Item58	1.160	0.062	-	-	1.216	0.047	-	-	0.613	-1.112	-	-
Item59	1.046	-0.225	-	-	1.288	0.051	-	-	0.863	-1.154	-	-
Item60	1.192	-0.190	-	-	1.562	-0.462	-	-	1.134	0.063	-	-

	English				Afrikaans				Sepedi			
	a	d	d1	d2	a	d	d1	d2	a	d	d1	d2
Item61	1.934	1.195	-	-	2.151	-0.075	-	-	1.015	-0.542	-	-
Item62	1.746	0.888	-	-	1.751	0.338	-	-	0.809	-0.841	-	-
Item63	2.344	1.412	-	-	1.451	0.417	-	-	1.389	-0.597	-	-
Item64	1.469	0.501	-	-	1.767	0.397	-	-	0.721	-0.428	-	-
Item65	1.007	-1.462	-	-	0.764	-0.815	-	-	0.360	-1.724	-	-
Item66	0.688	-0.226	-	-	0.434	-0.661	-	-	0.061	-0.600	-	-
Item67	2.199	1.236	-	-	1.873	0.207	-	-	1.052	-0.657	-	-
Item68	1.585	3.385	-	-	2.608	2.023	-	-	1.658	1.765	-	-
Item69	1.087	-0.337	-	-	0.804	-0.717	-	-	0.493	-0.684	-	-
Item70	0.615	-0.481	-	-	0.610	-0.594	-	-	0.337	-0.585	-	-
Item71	1.364	0.229	-	-	0.847	-0.542	-	-	0.519	-0.718	-	-
Item72	1.245	-0.361	-	-	1.010	-1.171	-	-	0.471	-1.230	-	-
Item73	2.341	3.550	-	-	2.491	2.711	-	-	1.771	1.693	-	-
Item74	2.594	2.300	-	-	1.414	0.383	-	-	0.625	-0.135	-	-
Item75	0.912	-0.319	-	-	1.464	-0.328	-	-	0.342	-1.038	-	-
Item76	1.071	0.172	-	-	1.468	-0.080	-	-	0.586	-0.747	-	-
Item77	1.074	-0.347	-	-	0.985	-0.440	-	-	1.142	-0.401	-	-
Item78	1.998	0.831	-	-	1.643	0.246	-	-	0.565	-0.941	-	-
Item79	0.467	-0.681	-	-	0.691	-0.570	-	-	0.442	-0.725	-	-
Item80	3.036	0.871	-	-	1.894	0.751	-	-	0.879	-0.674	-	-
Item81	0.879	0.184	-	-	1.118	-0.090	-	-	0.294	-0.804	-	-

	English				Afrikaans				Sepedi			
	a	d	d1	d2	a	d	d1	d2	a	d	d1	d2
Item82	1.664	0.685	-	-	2.249	1.475	-	-	1.454	0.465	-	-
Item83	1.443	0.136	-	-	1.320	-0.513	-	-	0.773	-0.505	-	-
Item84	1.341	-0.222	-	-	1.002	-0.510	-	-	-0.188	-1.118	-	-
Item85	0.513	-0.879	-	-	-0.027	-0.945	-	-	-0.592	-1.306	-	-
Item86	0.276	-0.140	-	-	0.459	-0.248	-	-	0.514	-0.514	-	-
Item87	0.854	-1.146	-	-	0.166	-0.598	-	-	0.202	-1.206	-	-
Item88	0.856	-1.204	-	-	0.653	-1.587	-	-	0.225	-1.801	-	-
Item89	0.788	-0.694	-	-	0.070	-0.899	-	-	0.230	-1.275	-	-
Item90	0.330	-1.080	-	-	0.103	-0.697	-	-	-0.174	-1.098	-	-
Item91	0.693	1.171	-	-	1.001	0.642	-	-	2.038	0.462	-	-
Item92	1.569	0.557	-	-	1.675	-0.531	-	-	0.854	-1.648	-	-
Item93	1.582	1.030	-	-	2.461	0.915	-	-	1.626	-0.402	-	-
Item94	2.032	2.156	-	-	2.567	1.629	-	-	2.138	0.521	-	-
Item95	0.768	0.470	-	-	1.508	0.453	-	-	-0.058	-1.157	-	-
Item96	1.928	0.102	-	-	1.765	-0.284	-	-	1.141	-1.586	-	-
Item97	2.050	-1.352	-	-	2.117	-1.544	-	-	0.808	-2.345	-	-
Item98	1.404	-0.616	-	-	0.818	-1.237	-	-	0.820	-1.587	-	-
Item99	2.227	0.523	-	-	2.080	-0.001	-	-	2.105	-2.417	-	-
Item100	1.663	-0.182	-	-	2.811	0.785	-	-	1.772	-1.624	-	-
Item101	3.018	0.418	-	-	3.038	0.798	-	-	1.966	-1.090	-	-
Item102	4.939	4.228	-	-	3.006	3.826	-	-	1.831	0.277	-	-

	English				Afrikaans				Sepedi			
	a	d	d1	d2	a	d	d1	d2	a	d	d1	d2
Item103	1.760	0.646	-	-	2.057	0.799	-	-	0.516	-0.856	-	-
Item104	2.032	0.398	-	-	2.142	0.897	-	-	2.035	-1.061	-	-
Item105	1.727	2.017	-	-	1.912	1.362	-	-	1.433	0.467	-	-
Item106	2.401	0.576	-	-	1.854	-0.032	-	-	2.135	-0.966	-	-
Item107	2.398	-0.030	-	-	1.652	-0.601	-	-	2.571	-1.348	-	-
Item108	1.627	-0.690	-	-	2.502	-0.362	-	-	1.308	-1.146	-	-
Item109	1.221	0.082	-	-	1.282	0.083	-	-	1.759	-0.894	-	-
Item111	1.428	0.738	-	-	3.067	2.340	-	-	0.993	0.052	-	-
Item112	1.222	0.079	-	-	1.979	0.205	-	-	1.795	-0.642	-	-
Item114	1.829	0.492	-	-	1.388	-0.099	-	-	0.489	-1.611	-	-
Item116	1.405	0.398	-	-	1.341	0.305	-	-	1.107	-1.559	-	-
Item117	1.364	0.264	-	-	2.008	0.242	-	-	1.258	-1.227	-	-
Item118	2.540	0.743	-	-	1.959	0.004	-	-	1.048	-0.033	-	-
Item119	1.563	-1.315	-	-	1.219	-0.865	-	-	1.026	-1.608	-	-
Item120	1.141	-1.845	-	-	0.762	-1.523	-	-	0.065	-2.192	-	-
Item123	0.506	-0.939	-	-	1.168	-0.542	-	-	1.665	-2.509	-	-
Item124	1.575	1.868	-	-	2.335	1.930	-	-	1.377	0.107	-	-
Item125	1.152	0.467	-	-	3.067	1.591	-	-	0.914	-1.453	-	-
Item126	1.113	-0.570	-	-	1.158	-0.286	-	-	1.553	-1.495	-	-
Item127	1.420	-0.684	-	-	1.688	-1.367	-	-	1.376	-1.212	-	-
Item128	1.921	0.677	-	-	1.866	-0.221	-	-	1.690	-1.115	-	-

	English				Afrikaans				Sepedi			
	a	**d**	**d1**	**d2**	**a**	**d**	**d1**	**d2**	**a**	**d**	**d1**	**d2**
Item129	0.641	0.643	-	-	1.835	0.923	-	-	2.379	1.232	-	-
Item130	1.375	2.075	-	-	2.689	2.662	-	-	1.547	-0.108	-	-
Item131	1.727	-0.852	-	-	1.486	-0.832	-	-	1.239	-1.678	-	-
Item132	1.098	-0.853	-	-	1.144	-0.795	-	-	1.434	-1.744	-	-
Item134	1.434	-0.233	-	-	1.379	-0.679	-	-	1.183	-1.291	-	-
Item135	2.739	3.078	-	-	3.541	3.808	-	-	1.451	1.598	-	-
Item136	1.572	0.645	-	-	0.958	-1.511	-	-	1.166	-0.493	-	-
Item137	2.355	2.030	-	-	2.769	1.641	-	-	1.725	0.576	-	-
Item138	2.968	1.672	-	-	1.738	0.135	-	-	1.862	-0.213	-	-
Item140	1.216	0.825	-	-	1.485	0.874	-	-	1.427	-0.370	-	-
Item141	1.312	-0.603	-	-	1.802	-1.533	-	-	1.349	-1.593	-	-
Item142	1.377	0.310	-	-	1.510	0.114	-	-	1.168	-1.049	-	-
Item143	2.254	0.461	-	-	1.958	-0.030	-	-	1.103	-1.345	-	-
Item144	3.210	2.605	-	-	4.412	2.708	-	-	3.223	0.366	-	-
Item145	3.331	1.801	-	-	2.965	0.352	-	-	2.109	-0.313	-	-
Item146	1.981	-0.743	-	-	1.769	-1.894	-	-	2.210	-2.997	-	-
Item147	1.648	-0.480	-	-	1.183	-0.930	-	-	0.830	-1.541	-	-
Item148	1.540	0.319	-	-	1.108	-0.150	-	-	0.809	-0.825	-	-
Item149	3.270	1.997	-	-	1.585	-0.353	-	-	2.115	-2.444	-	-
Item151	2.359	2.619	-	-	4.046	3.612	-	-	1.269	0.481	-	-
Item152	2.794	-0.194	-	-	3.316	-0.470	-	-	1.829	-1.817	-	-

	English				Afrikaans				Sepedi			
	a	**d**	**d1**	**d2**	**a**	**d**	**d1**	**d2**	**a**	**d**	**d1**	**d2**
Item153	2.135	1.183	-	-	2.896	0.651	-	-	1.513	-0.675	-	-
Item154	2.179	-0.318	-	-	2.427	-1.702	-	-	1.309	-2.717	-	-
Item155	2.287	-1.126	-	-	2.376	-1.919	-	-	1.406	-1.780	-	-
Item157	1.119	0.000	1.110	2.017	1.264	0.000	0.702	0.687	1.065	0.000	0.405	-0.153
Item158	1.592	-1.167	-	-	1.910	-1.450	-	-	1.248	-2.104	-	-
Item159	1.959	0.454	-	-	1.716	0.632	-	-	1.753	-1.585	-	-
Item160	1.061	-0.753	-	-	1.104	-0.730	-	-	1.337	-1.529	-	-
Item161	2.327	-0.606	-	-	1.757	-0.126	-	-	1.331	-1.924	-	-
Item162	1.103	0.000	0.185	-0.306	1.486	0.000	0.791	0.219	0.956	0.000	-0.518	-2.386
Item163	1.815	0.616	-	-	1.762	1.176	-	-	0.409	-0.815	-	-
Item164	1.357	0.000	-0.572	1.058	1.091	0.000	-0.514	0.318	0.736	0.000	-1.100	-0.775
Item165	1.969	0.536	-	-	1.274	0.087	-	-	1.067	-0.739	-	-
Item168	1.429	0.286	-	-	2.523	0.201	-	-	0.933	-1.572	-	-
Item169	1.018	0.000	-0.393	-0.555	1.454	0.000	-0.500	-0.576	1.154	0.000	-1.880	-2.094
Item170	0.888	0.000	-0.881	-0.642	1.180	0.000	-0.210	-0.919	1.078	0.000	0.075	-1.699
Item171	0.873	-0.604	-	-	1.065	-0.503	-	-	1.213	-1.815	-	-
Item172	1.980	0.000	1.184	3.254	2.115	0.000	1.255	2.395	1.053	0.000	0.164	0.090
Item173	2.217	0.623	-	-	2.209	0.003	-	-	0.564	-1.181	-	-
Item174	0.645	0.000	-0.085	-0.245	0.678	0.000	0.473	-0.427	0.409	0.000	-0.966	-1.879
Item175	3.026	1.125	-	-	2.677	-0.099	-	-	1.870	-0.624	-	-
Item176	1.636	-2.549	-	-	1.358	-2.194	-	-	1.297	-2.595	-	-

	English				Afrikaans				Sepedi			
	a	**d**	**d1**	**d2**	**a**	**d**	**d1**	**d2**	**a**	**d**	**d1**	**d2**
Item177	1.763	0.289	-	-	1.979	0.474	-	-	1.402	-1.225	-	-
Item180	1.595	0.000	0.396	0.009	1.508	0.000	0.282	-0.234	0.896	0.000	-0.283	-1.651
Item181	0.965	-1.234	-	-	1.164	-2.229	-	-	0.835	-2.359	-	-
Item182	3.708	1.412	-	-	4.497	-0.179	-	-	2.597	-0.929	-	-

Appendix E Paper 3 additional information

E-1 Passage-level differential response functioning statistics

Table A-16

Differential response functioning statistics at the passage level – Afrikaans focal group

Passage	Focal group	Number of items	Statistic	DRF	5% CI	95% CI	χ^2	df	p-value
A	Afrikaans	sDRF	14	-0.258	-0.722	0.129	1.441	1	0.230
A	Afrikaans	uDRF	14	0.262	0.090	0.722	4.119	2	0.128
B	Afrikaans	sDRF	15	0.623	0.212	0.989	9.437	1	0.002
B	Afrikaans	uDRF	15	0.632	0.276	1.000	11.643	2	0.003
C	Afrikaans	sDRF	13	0.465	0.100	0.844	5.863	1	0.015
C	Afrikaans	uDRF	13	0.707	0.386	1.020	21.347	2	< 0.001
D	Afrikaans	sDRF	12	0.505	0.088	0.923	5.762	1	0.016
D	Afrikaans	uDRF	12	0.514	0.209	0.924	8.742	2	0.013
E	Afrikaans	sDRF	15	0.667	0.333	0.997	15.685	1	< 0.001
E	Afrikaans	uDRF	15	0.671	0.367	1.003	19.338	2	< 0.001
f	Afrikaans	sDRF	13	0.166	-0.591	0.879	0.203	1	0.653
f	Afrikaans	uDRF	13	0.259	0.086	0.923	1.650	2	0.438
G	Afrikaans	sDRF	13	-0.571	-0.896	-0.224	11.924	1	0.001
G	Afrikaans	uDRF	13	0.571	0.244	0.885	24.681	2	< 0.001
H	Afrikaans	sDRF	15	0.194	-0.223	0.622	0.862	1	0.353
H	Afrikaans	uDRF	15	0.214	0.074	0.631	2.350	2	0.309
I	Afrikaans	sDRF	12	0.255	-0.415	0.862	0.615	1	0.433
I	Afrikaans	uDRF	12	0.269	0.084	0.889	1.943	2	0.378
J	Afrikaans	sDRF	18	-0.215	-0.600	0.177	1.146	1	0.284
J	Afrikaans	uDRF	18	0.245	0.107	0.617	6.831	2	0.033
K	Afrikaans	sDRF	8	0.007	-0.248	0.280	0.002	1	0.960
K	Afrikaans	uDRF	8	0.034	0.043	0.331	0.328	2	0.849
L	Afrikaans	sDRF	4	0.025	-0.101	0.158	0.151	1	0.697
L	Afrikaans	uDRF	4	0.138	0.057	0.265	8.367	2	0.015

Table A-17

Differential response functioning statistics at the passage level – Sepedi focal group

Passage	Focal group	Number of items	Statistic	DRF	5% CI	95% CI	χ^2	df	p-value
A	Sepedi	sDRF	14	0.160	-0.365	0.599	0.424	1	0.515
A	Sepedi	uDRF	14	0.432	0.171	0.876	6.751	2	0.034
B	Sepedi	sDRF	15	0.155	-0.334	0.595	0.429	1	0.512
B	Sepedi	uDRF	15	0.844	0.514	1.255	26.996	2	< 0.001
C	Sepedi	sDRF	13	0.560	0.045	0.982	5.166	1	0.023
C	Sepedi	uDRF	13	0.709	0.234	1.127	11.712	2	0.003
D	Sepedi	sDRF	12	-0.198	-0.737	0.264	0.610	1	0.435
D	Sepedi	uDRF	12	0.549	0.271	1.038	11.286	2	0.004
E	Sepedi	sDRF	15	0.248	-0.223	0.631	1.197	1	0.274
E	Sepedi	uDRF	15	0.485	0.293	0.841	14.511	2	0.001
f	Sepedi	sDRF	13	0.003	-0.712	0.591	0.000	1	0.993
f	Sepedi	uDRF	13	0.824	0.470	1.356	16.778	2	< 0.001
G	Sepedi	sDRF	13	0.662	0.239	1.010	10.843	1	0.001
G	Sepedi	uDRF	13	0.664	0.312	1.021	13.223	2	0.001
H	Sepedi	sDRF	15	-1.127	-1.625	-0.666	21.626	1	< 0.001
H	Sepedi	uDRF	15	1.291	0.864	1.794	56.236	2	< 0.001
I	Sepedi	sDRF	12	-0.258	-0.948	0.340	0.621	1	0.431
I	Sepedi	uDRF	12	0.675	0.373	1.214	14.172	2	0.001
J	Sepedi	sDRF	18	-0.893	-1.346	-0.448	15.579	1	< 0.001
J	Sepedi	uDRF	18	0.898	0.558	1.375	42.336	2	< 0.001
K	Sepedi	sDRF	8	-0.674	-0.948	-0.372	20.977	1	< 0.001
K	Sepedi	uDRF	8	0.727	0.449	1.025	48.101	2	< 0.001
L	Sepedi	sDRF	4	0.120	-0.040	0.303	1.910	1	0.167
L	Sepedi	uDRF	4	0.213	0.108	0.354	13.208	2	0.001