A revised PIRLS 2011 to 2016 trend for South Africa and the importance of analysing the underlying microdata

MARTIN GUSTAFSSON

Stellenbosch Economic Working Papers: WP02/2020

www.ekon.sun.ac.za/wpapers/2020/wp022020

February 2020

KEYWORDS: South Africa, PIRLS, international testing systems, reading
JEL: C13, C89, I21

ReSEP (Research on Socio-Economic Policy)
http://resep.sun.ac.za

DEPARTMENT OF ECONOMICS
UNIVERSITY OF STELLENBOSCH

SOUTH AFRICA

UNIVERSITEIT STELLENBOSCH UNIVERSITY

BER
BUREAU FOR ECONOMIC RESEARCH

A WORKING PAPER OF THE DEPARTMENT OF ECONOMICS AND THE
BUREAU FOR ECONOMIC RESEARCH AT THE UNIVERSITY OF STELLENBOSCH

www.ekon.sun.ac.za/wpapers

A revised PIRLS 2011 to 2016 trend for South Africa and the importance of analysing the underlying microdata

MARTIN GUSTAFSSON

JANUARY 2020

ABSTRACT

Given South Africa's weak performance in international testing programmes, there is a strong interest in gauging improvements within these programmes. The finding that South Africa saw no progress between 2011 and 2016 in the Progress in International Reading Literacy Study (PIRLS) programme, which tests Grade 4 reading competencies, was inconsistent with considerable progress seen in a couple of other testing programmes. Moreover, an earlier PIRLS average score for Grade 4 from 2006 suggested that the 2011 mean score used to determine the flat 2011 to 2016 flat trend was problematic. The current paper uses the underlying microdata for PIRLS 2011 and 2016, which are publicly available, to examine the trend. It is clear that the 2011 mean score used by the international PIRLS analysts to arrive at the flat trend cannot be correct. It should be considerably lower. It should be noted that the 2011 mean for South Africa involved an unusual process. South Africa was the only country for which an original mean on an easier scale, prePIRLS, had to be recalibrated to the main PIRLS scale. This was because South Africa was the only country participating in the easier prePIRLS in 2011 and in some form of PIRLS in 2016. There was clearly something wrong with the 2011 recalibration. In correspondence, the International Association for the Evaluation of Educational Achievement (IEA), the body that conducts PIRLS globally, acknowledged, in part on the basis of a preliminary version of the current paper, that the originally published 2011 to 2016 South Africa trend should now not be considered reliable. They also confirmed that the classical score gains for South Africa reflected in the paper are correct. The method used in the paper is essentially to examine classical score gains between 2011 and 2016 with respect to common items, and then to recalibrate that to the main PIRLS scale. As an additional verification, the paper checks that gains remain after one controls for socio-economic status. The paper concludes that South Africa in fact saw a large gain between 2011 and 2016, equal to around 0.05 standard deviations a year. This is a fast rate of improvement by international standards. Of the 43 countries with 2011 to 2016 trends in PIRLS, South Africa displayed the third-steepest improvement, after Morocco and Oman.

Martin Gustafsson
Department of Economics
University of Stellenbosch
Private bag X1, 7602
Matieland, South Africa
E-mail: mgustafsson@sun.ac.za

RESEP
Research on Socio-Economic Policy

The author is based part-time at the Department of Basic Education in Pretoria, and is Associate Professor of Economics at the University of Stellenbosch.

1 Introduction

The current paper reports on the results of an analysis of the publicly available data from the Progress in International Reading Literacy Study (PIRLS) programme. The analysis was undertaken to see whether the official 2011 to 2016 PIRLS trend for South Africa, which was a flat no-change trend, was indeed correct. Valid concerns had been raised that a flat trend seemed incorrect, given what a wider set of available information both within PIRLS and beyond PIRLS was saying. It should be underlined that the official 2011 to 2016 PIRLS trend for South Africa was based on a 2011 value which was the outcome of a recalibration process applied only to South Africa. Concerns and uncertainties rested in part on a lack of clarity around this process.

The main question the paper addresses is thus what the trend in the reading abilities of children in the national sample between 2011 and 2016 really was. Section 2 explains the PIRLS trends for South Africa reported to date. Section 3 summarises what exists in the public domain on the methodology used by PIRLS analysts to recalibrate South Africa's 2011 PIRLS results. Section 4 explains the structure of the PIRLS tests and data, or at least those aspects which are relevant for the paper. Section 5 examines the 2011 to 2016 trends for South Africa, using the raw item-level data, and provides an estimate of what the 2011 national score should be. Section 6 analyses a question which is not integral to the paper's main question, yet is important for understanding the trend: were the two samples, of 2011 and 2016, sufficiently representative and comparable? Section 7 concludes.

The analysis presented below was initially produced as part of the author's work for the South African Department of Basic Education (DBE), the national authority responsible for schools. It was discussed with relevant analysts in South Africa and those working for the IEA[1], the organisation that conducts PIRLS globally. IEA analysts agreed, in part on the basis of a preliminary version of the current paper, that the originally published 2011 to 2016 South Africa trend should now not be considered reliable. They also confirmed that the classical score gains for South Africa described below are correct.

While the current paper raises, implicitly, important questions around future validation processes in testing programmes such as PIRLS and, crucially, data analysis capacity at the country level, it is not the intention to explore those questions here. They obviously need to be explored elsewhere. Similarly, the paper does not speculate on what a revised trend means for understanding the recent history of the South African schooling system, for instance in terms of what interventions or factors may be responsible for the trend.

2 Publicly available PIRLS trend information for South Africa

In 2017, the international report of the 2016 wave of the Progress in International Reading Literacy Study (PIRLS) programme was published. Also in 2017, the Centre for Evaluation and Assessment (CEA), the local implementer of PIRLS 2016 in South Africa, based at the University of Pretoria, released its own report. Both reports pointed to there having been no improvement in reading in South Africa between 2011 and 2016. The average score for Grade 4 reading in South Africa was said to be 323 in 2011 and 320 in 2016. These statistics, and their source, appear in Table 1 below. This trend essentially represents no change after margins of error in the sampled-based statistics are considered. The national report by CEA simply replicated the findings produced by Boston College, the institution which has historically analysed data for the IEA. The CEA did not conduct its own analysis of the trend, using the raw data it had access to.

[1] International Association for the Evaluation of Educational Achievement.

Table 1: 2006, 2011 and 2016 national Grade 4 values

PIRLS year	National Grade 4 score	Where published	Comment
2006	253	Howie *et al*, 2008: 19.	A national Grade 5 score of 302 was published in Mullis *et al* (2007: 37)
2011	323	Mullis *et* al, 2017: 29	The 2011 national score using the prePIRLS scale is 461 and appears in Mullis *et al* (2012: 39).
2016	320	Mullis *et* al, 2017: 29	

2017 was the first year in which an official PIRLS trend for South Africa was pronounced. The 2011 score of 323 was the result of a recalibration of an original value of 461 expressed in terms of the easier prePIRLS scale employed across some countries only in 2011. In 2016, a different test for developing countries with low performance was introduced, PIRLS Literacy, but this test produced scores comparable to the main PIRLS tests, because there were items shared across the two that allowed equating of results across the tests. In 2016, South Africa, together with Egypt, Iran, Kuwait and Morocco, took the PIRLS Literacy tests.

One thing that raised concerns around the correctness of the flat 2011 to 2016 trend was the 2006 Grade 4 score of 253. In 2006, though South Africa officially entered Grade 5 in PIRLS – this was within the main PIRLS system as no easier PIRLS existed – South Africa also tested a sample of Grade 4 learners. In its 2017 report, the CEA indicated, correctly, that the difficulty of the tests used in 2006 was such that the Grade 4 score of 253 was highly unreliable. The CEA has therefore always left this 2006 score out of its discussion of trends[2]. After the publication of the flat 2011 to 2016 trend, the author of the current paper and others working for the DBE at the time argued that even if the 2006 mean score was unreliable, it should not be completely ignored, especially given that it produced a mysterious pattern. If the 2006 score is used, the 2006 to 2011 gain, of 70 points (323 minus 253), is twice as large as the largest 2006 to 2011 national gain published in the 2011 international report, namely that seen in Trinidad and Tobago[3]. For such an exceptional gain to be followed by a flat 2011 to 2016 trend raised questions around the 2011 score of 323. It seemed reasonable to assume that the 2011 score should be lower, producing a more consistent trend over the ten years 2006 to 2016. The difference between the 2006 Grade 5 score of 302, a score which everyone agreed was reliable, and the 2006 Grade 4 score of 253, a difference of 49 points, or half a PIRLS standard deviation, was believable, which seemed to strengthen the argument that the 253 value was at least roughly reliable. A separate South African dataset, that of the National School Effectiveness Study (NSES), points to grade-on-grade gains in reading of 0.49 of a standard deviation a year[4].

Apart from testing a national sample, in 2006, 2011 and 2016 South Africa also tested samples of children which were not nationally representative, but representative only of specific language sub-groups in the country. Results for these non-national samples have been reported on in various places, including the international reports. The current report does not analyse trends in these sub-samples, as this is not directly necessary to answer the main question of the paper, which is to establish the veracity of the flat 2011 to 2016 trend for the country. However, the sub-samples are clearly important for understanding reading trends in South Africa in a broader sense.

[2] Howie *et al*, 2017: 82.
[3] Mullis *et al*, 2012: 50.
[4] From Gustafsson, Mabogoane and Taylor (2012: 23) and Taylor and Taylor (2013: 16).

3 How the 2011 IRT score for South Africa was recalibrated

There is little publicly available on how Boston College converted South Africa's 2011 prePIRLS score to the main PIRLS scale. However, the following appears in a PIRLS 2016 technical report[5]:

> In 2011, PIRLS Literacy's predecessor prePIRLS was reported as its own scale, although its item parameters were estimated on the same item parameter metric, capitalizing on Colombia's participation in both PIRLS and prePIRLS in 2011. However, with South Africa having participated in both prePIRLS in 2011 and PIRLS Literacy in 2016, there was a need to place their 2011 results on the PIRLS trend scale. To that end, it was necessary to re-transform their achievement scores – overall reading, as well as the purposes and processes – using the PIRLS 2011 linear transformation constants…

By not mentioning South Africa's 2016 PIRLS Literacy data, or the 2016 item parameters, this extract could be taken to mean that just 2011 data, and specifically Colombia's 2011 data, were used as a basis for recalibrating South Africa's 2011 prePIRLS values. If that is the case, then a few questions arise. In particular, how were the Colombia data used to produce a recalibration, given that there were no common items across prePIRLS 2011 and main PIRLS 2011? If the assumption was that the two Colombia samples in 2011, one for regular PIRLS and one for prePIRLS, reflected the same population, then what risk is there that this assumption did not hold, meaning that the two samples were substantively dissimilar? Given that Colombia performs considerably better than South Africa – in 2016 surpassing South Africa by 128 PIRLS points, on the main PIRLS scale[6] – how applicable would a conversion algorithm using just Colombia data be for South Africa?

4 The structure of the PIRLS tests and data

Very fortunately, PIRLS (and TIMSS) data are exceptionally well documented, and the microdata are downloadable off the PIRLS-TIMSS website (timssandpirls.bc.edu). Without this, analyses of the kind produced below would not be possible. The PIRLS-TIMSS website can be considered a 'gold standard' for other testing systems, including national ones.

The values 323 and 320, for the years 2011 and 2016, referred to in Table 1, were successfully replicated, using the downloaded microdata. The analysis used the command 'pv' in Stata 16[7]. This provided certainty that the right data were being used. The 2016 data contained the results of 12,810 test-takers in 293 schools. The median number of test-takers per school was 47, with 27 at the 5th percentile and 86 at the 95th percentile (the pupil weights in the data were used in calculating this). In 2011, there were 15,744 test-takers in 341 schools, the median number of test-takers in a school being 50 (26 at the 5th percentile, 111 at the 95th percentile).

Table 2 below reflects the structure of the 2016 PIRLS Literacy tests. Each test booklet contained two passages, plus questions relating to each passage. Of the twelve passages, four were also used in the main PIRLS 2016 tests, and another four were used in prePIRLS 2011. It is the data from these latter four passages which are used to establish the actual South African trend between 2011 and 2016.

Table 3 below provides basic details on the performance of the 2011 and 2016 samples. The statistics in this table are not weighted. The four additional passages, in category 'D' (this

[5] Martin *et al*, 2017: 12.18.

[6] Colombia's mean in 2011 was 448 (on the main PIRLS scale), compared to South Africa's 320 in 2016. In 2016, Colombia did not participate in PIRLS.

[7] The full command line for both years was 'pv, pv(ASRREA0*) jkzone(JKZONE) jkrep(JKREP) weight(TOTWGT) jrr timss: mean @pv [aw=@w]'.

categorisation is my own), are four passages used in prePIRLS 2011, and nowhere else. The items for the four passages used in both 2011 and 2016 come to a total of 65 items. Of the 65, 31 were four-option multiple choice (MC) questions, and 34 constructed response (CR) questions. The maximum score per passage mostly exceeds the sum of the two types of items because certain constructed response items carry a maximum score of not one, but two or even three points. The percentage of pupils with a score of zero, and the mean classical scores (or percentage scores) make it clear that pupils scored higher in the multiple choice questions. This would be due to the possibility of random guessing. Mean score values are means across all pupils, including those with zero or missing. A missing response was counted as zero throughout the table. If one combines MR and CR, 2.5% of pupils in 2016 and 5.7% in 2011 achieved a score of zero, counting both passages.

Table 2: South Africa counts for PIRLS Literacy passages and books in 2016

Cat.	Passage	L01	L02	L03	L04	L05	L06	L07	L08	L09	L10	L11	L12	L13	L14	L15	LR	Pupils
A	Flowers on the Roof	X								X				X				2,091
A	How Did We Learn to Fly?			X		X							X					2,135
A	Pemba Sherpa								X		X				X			2,168
A	Sharks				X		X					X						2,107
B	Ants									X	X					X		2,135
B	Baghita's Perfect Orange		X	X								X						2,131
B	The Summer My Father Was 10				X	X										X		2,125
B	Training A Deaf Polar Bear							X	X					X				2,132
C	African Rhinos																X	2,113
C	Hungry Plant	X	X												X			2,108
C	Library Mouse						X	X					X					2,105
C	The Pearl																X	2,132
	Pupils	694	710	709	704	723	703	710	726	711	725	720	704	708	720	710	2,133	12,810

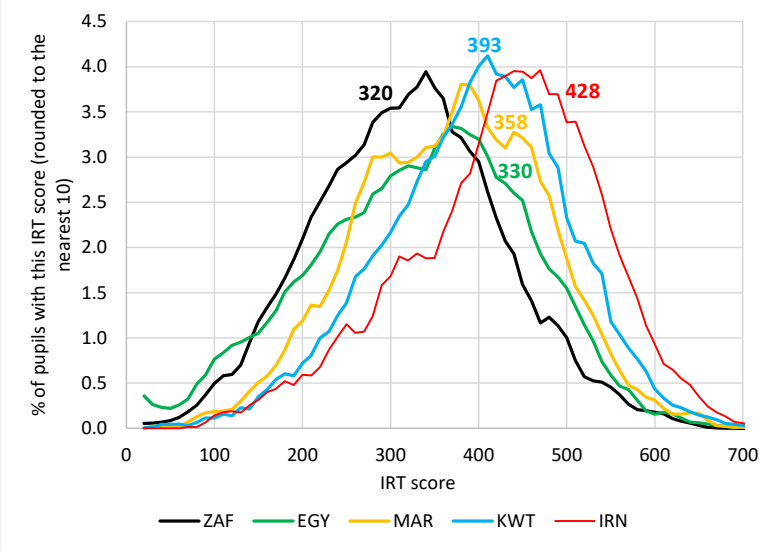
Note: In the first column, A refers to passages shared with regular PIRLS in 2016, B refers to passages shared with prePIRLS 2011, and C refers to PIRLS Literacy passages not shared elsewhere.

Table 3: PIRLS South Africa details per passage for 2011 and 2016

		Pupils		Items			% with zero MC		% with zero CR		Mean score MC		Mean score CR	
Cat.		2011	2016	MC	CR	Max. score	2011	2016	2011	2016	2011	2016	2011	2016
A	Flowers on the Roof		2,091	7	6	16		14		48		31		12
A	How Did We Learn to Fly?		2,135	8	9	19		6		9		37		38
A	Pemba Sherpa		2,168	8	9	20		4		16		49		27
A	Sharks		2,107	6	6	17		19		41		26		13
B	Ants	3,849	2,135	9	12	25	8	4	22	10	39	49	24	35
B	Baghita's Perfect Orange	3,873	2,131	8	8	17	11	3	25	14	41	53	27	36
B	The Summer My Father Was 10	3,870	2,125	8	6	15	12	5	24	14	39	48	31	38
B	Training A Deaf Polar Bear	3,853	2,132	6	8	16	21	11	25	14	34	42	24	32
C	African Rhinos		2,113	7	10	19		8		11		44		40
C	Hungry Plant		2,108	11	5	16		3		16		42		43
C	Library Mouse		2,105	9	9	21		3		7		47		42
C	The Pearl		2,132	7	8	18		8		13		44		36
D	Brave Charlotte	3,826		5	12	20	22		20		34		28	
D	Caterpillar to B	3,894		8	8	16	16		18		34		35	
D	Lonely Giraffe	3,824		8	6	15	15		31		35		33	
D	Two Giant Dinosa	3,904		7	9	17	20		29		29		27	
	% with zero in both passages						10	3	7	2				

Turning to what will be referred to as IRT scores (IRT being item response theory), Figure 1 illustrates distributions for the five countries participating in PIRLS Literacy 2016. Pupil weights are used here. Here and in the remainder of the analysis, the first of the five plausible values available in the data was used as each pupil's IRT score. Had one of the other four been used, the picture would have looked essentially identical. Assuming there was no improvement between 2011 and 2016 in South Africa, one could assume that the 2011 distribution, using the main PIRLS scale, would look like the 2016 distribution. There are no publicly available pupil-level scores for 2011 using the main PIRLS scale. What are available, in the 2011 dataset, are pupil-level scores for South Africa using the easier prePIRLS scale. A key question for this paper is what the 2011 distribution would look like in Figure 1, if one recalibrated the 2011 prePIRLS values to the main PIRLS scale.

Figure 1: PIRLS Literacy IRT score distributions



Note: Apart from rounding at the level of pupils, to the nearest 10, curves are smoothed using simple moving averages across three values.

Figure 2 below illustrates the distribution of weighted classical scores for just the passage "Baghita's Perfect Orange". Each pupil's score was rounded to the closest multiple of 5. Here the 2011 distribution for South Africa can be included. As one would expect, given what appears in Table 2, the 2011 South Africa curve lies to the left of the 2016 South Africa curve. Similar patterns would be seen if similar graphs were generated for the other three passages shared between 2011 prePIRLS and 2016 PIRLS Literacy.

Figure 2: Histogram for 'Baghita…' classical scores

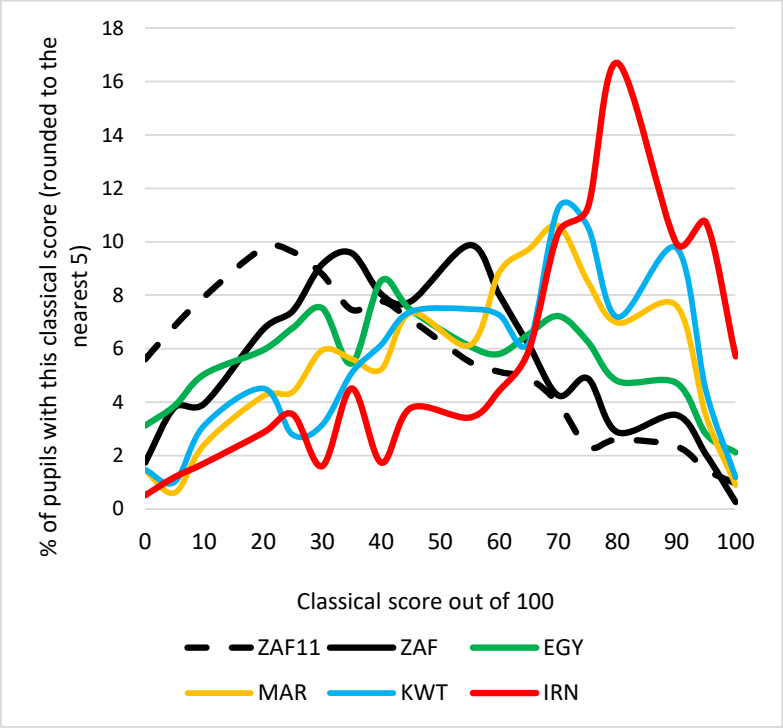


Figure 3 reflects the same distributions, but cumulatively. Here the better performance for this passage in 2011 compared to 2016 in the case of South Africa is even clearer.

Figure 3: Cumulative distribution for 'Baghita…' classical scores

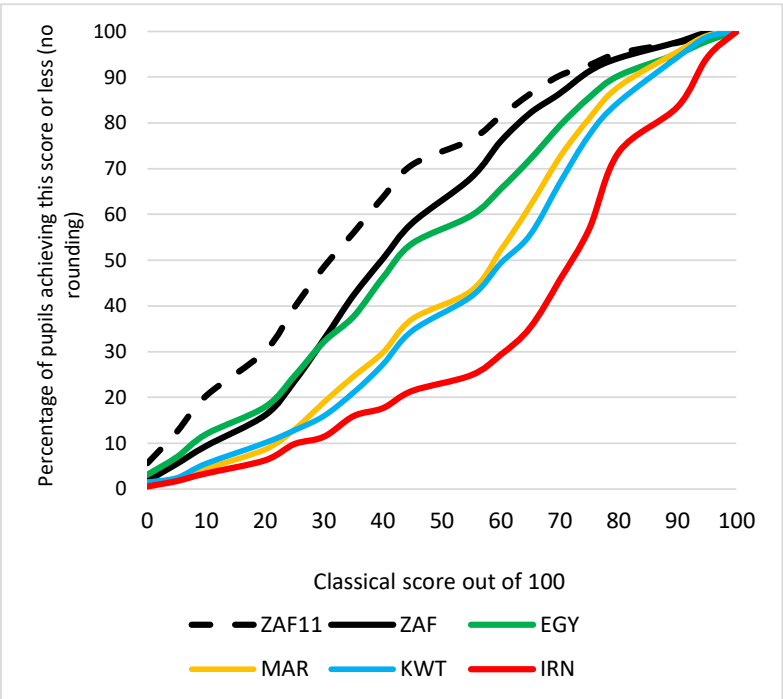


Table 5 provides weighted means of the classical scores for the four passages in question, and for five countries. For 2016, countries are arranged in an ascending order of performance.

Roughly, between 2011 and 2016, it appears South Africa improved by a margin equal to the gap between two countries, in particular the six-point gap between Morocco and Kuwait. This table suggests strongly that the true South Africa 2011 distribution in Figure 1 would lie noticeably to the left of South Africa's 2016 curve.

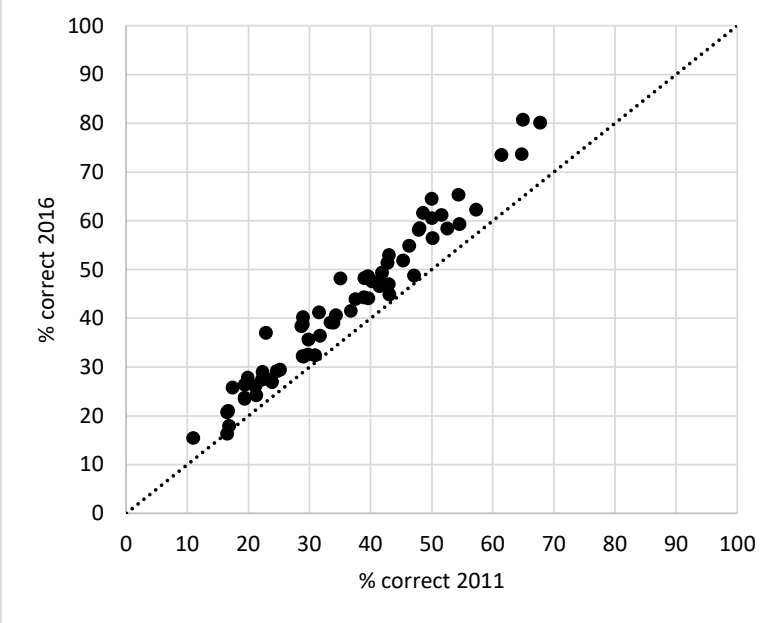
Table 4: Weighted classical scores in four passages, five countries

	2011	2016				
	South Africa	South Africa	Egypt	Morocco	Kuwait	Iran
Ants	33.7	41.4	49.5	51.6	63.8	66.2
Baghita's Perfect Orange	37.3	45.0	48.2	56.8	58.8	68.7
Training A Deaf Polar Bear	31.8	36.9	41.6	50.9	53.5	66.6
The Summer My Father Was 10	38.3	43.8	44.9	52.9	60.4	64.8
Mean across four passages	35.3	41.8	46.1	53.1	59.1	66.6

Note: Means in the bottom row are means across the four values in the column.

Figure 4 represents the gain seen in South Africa for each of the 65 common items. Importantly, the distribution suggests there was a general increase across all items. The 2011 to 2016 gains seen in the above table are not driven by large gains in just a few outlier items.

Figure 4: 2011-2016 gains by item



5 Estimating a new 2011 IRT score for South Africa

The aim in this section is not to replicate exactly what Boston College might do to produce a more accurate 2011 distribution of IRT scores, expressed in terms of the main PIRLS scale. The IRT methods employed by Boston College are readily available, for instance in Martin *et al* (2017: chapters 11 and 12). However, the variety of computer software they use to implement the methods are not readily available. In theory, it would be possible to code the methods in, for instance, Stata's Mata language. However, this would be a huge undertaking. The approach taken below is a relatively crude shortcut approach aimed at approximating, with a reasonable degree of accuracy, a 2011 IRT distribution, along the main PIRLS scale, that would faithfully reflect the gains over time clearly seen in the classical scores. One advantage with such an approach is that it helps to illustrate to the average reader the relationship between the classical and IRT scoring systems.

The functionality of the 'irt' command in Stata 16 'irt' is clearly better than that in previous versions of Stata. The command in Stata 16 is able to perform many of the required computations, though there were problems with some. One thing which did not appear possible, was to perform IRT computations using non-binary item scores, specifically the constructed response items carrying a maximum score of more than one point. The Stata 16 manuals provide instructions for doing this, but this always returned an error. Non-binary items were therefore excluded. Of the 65 common items across the four passages, 58 are said to be 'used for scaling' in the PIRLS documentation, and hence carry a 'slope' and a 'location' in the documentation. Stata uses the terms 'discrimination parameter' and 'difficulty parameter' to refer to the same things. Of the 58 items, 51 were binary.

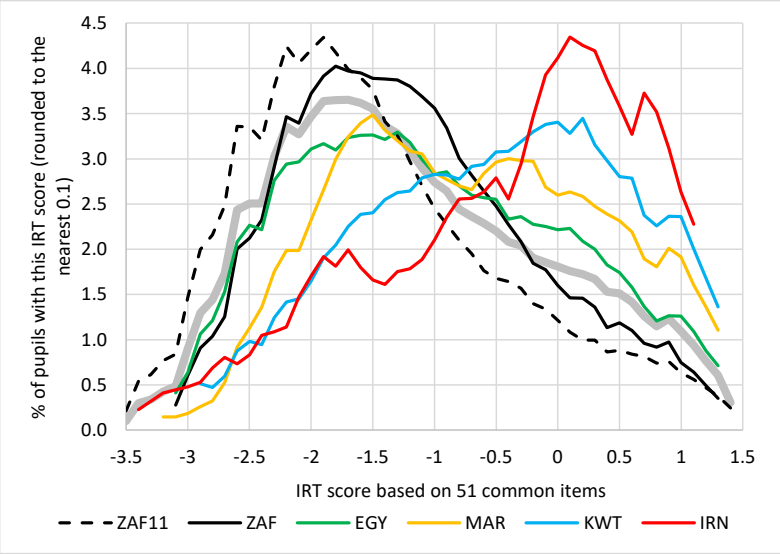
Table 5: Items used for IRT scoring

	Original items per passage	Used for scaling	After removing non-binary CR
Ants	21	14	11
Baghita's Perfect Orange	16	16	15
The Summer My Father Was 10	14	14	13
Training A Deaf Polar Bear	14	14	12
Total	65	58	51

The IRT computation had to use only pupils who responded to at least one of the four common passages. In the case of South Africa in 2016, this meant 61% of pupils could be included in the computation – this can be calculated from Table 2. A very similar percentage of included observations applied for the other four countries. In the case of the 2011 South Africa data, 83% of pupils responded to at least one of the four passages, and were therefore included in the computation.

All six samples, the five 2016 samples and South Africa's 2011 sample, were run through one single IRT computation. Discrimination and difficulty parameters were set for all 51 items, using the published PIRLS 2016 values. The distributions of the resulting pupil scores are illustrated in Figure 5, which uses pupil weights. The two South Africa curves are remarkably similar to each other, though the 2011 distribution is clearly to the left of the 2016 distribution, implying a considerable improvement in performance between 2011 and 2016.

Figure 5: IRT distributions based on 51 common items



Note: The grey curve is the distribution for the pooled six samples for 2016, where each country's pupil weights are recalibrated to produce an equal total across countries.

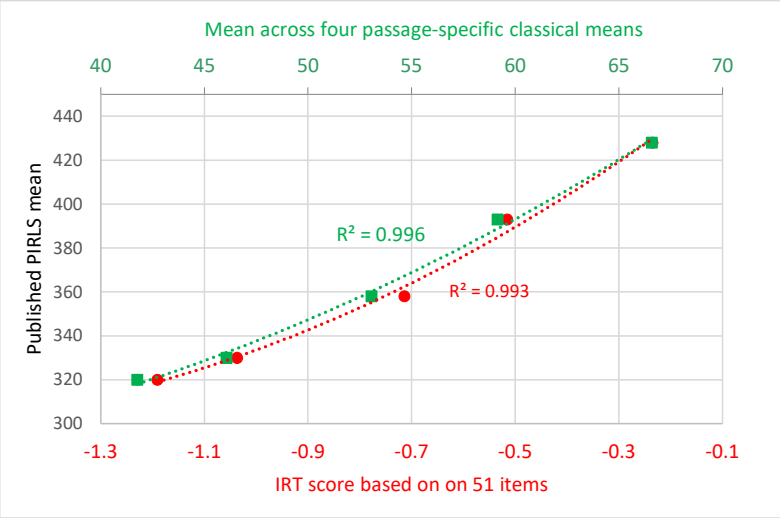
The unweighted standard deviation of the IRT scores in the pooled data covering all six samples was 1.11 (if weights are applied, it is 1.15). For just South Africa 2016 it is 0.94 and just South Africa 2011 it is 0.96, using no weights (with weights the values would be 0.98 and 1.05). The weighted means for South Africa are -1.48 in 2011 and -1.19 in 2016. This translates into an overall difference of 0.28 South African standard deviations, or 0.06 standard deviations a year. A gain of 0.06 standard deviations a year is about the highest annual gain one can expect, given historical trends across the world[8].

Once the comparable IRT scores had been computed, the next step was to find a method that would convert South Africa's IRT score based on just 51 items, to another IRT score that approximated the main PIRLS scale with all items included. Various approaches were tested. The one described below seemed the most transparent and credible.

The red markers in Figure 6 confirm that the IRT scores based on just 51 items predict the official and published 2016 means for the five PIRLS Literacy countries well. But the same can be said of the much simpler means across four passages (the bottom row of Table 5). The latter is illustrated by the green markers in Figure 6.

[8] UNESCO, 2019.

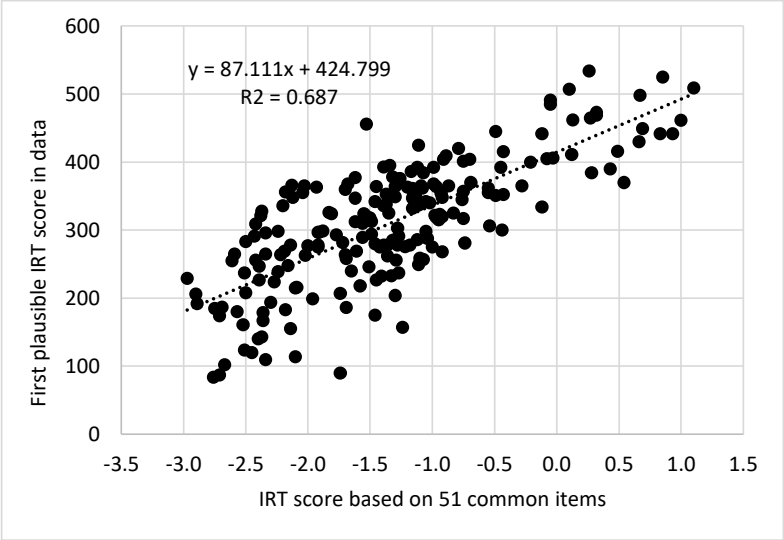
Figure 6: Predicting official means in five PIRLS Literacy countries



Note: Trendlines are quadratic.

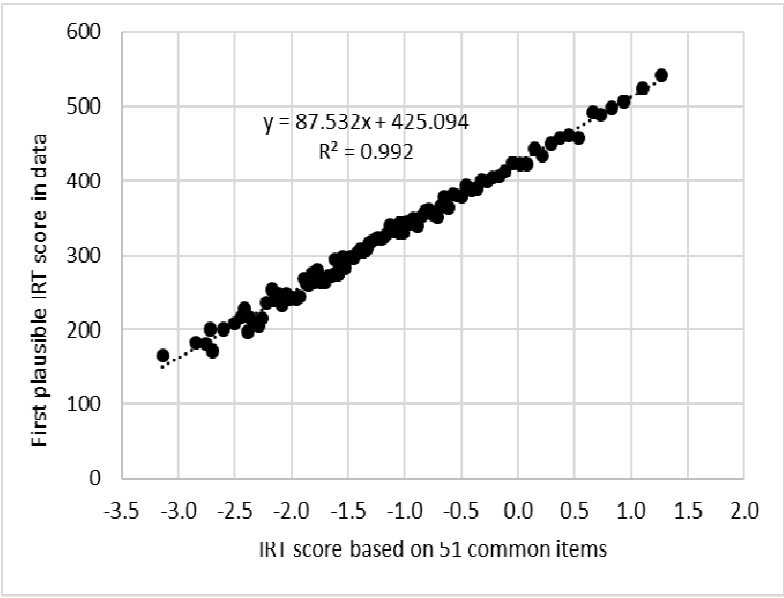
Figure 7, which uses just South Africa 2016 data, points to considerable 'noise' in the relationship between the two IRT scores at the level of the pupil. However, if weighted pupils in 2016 are aggregated into percentiles, based on their 51-item IRT score, the correlation is almost perfect – see Figure 8. It seemed optimal to estimate a new 2011 PIRLS mean using percentiles of the national sample.

Figure 7: Pupil IRT scores in South Africa's 2016 PIRLS Literacy



Note: Markers represent a random sub-sample of 200 from the larger 7,856 sample. Coefficients, however, reflect the regression run on all 7,856 observations.

Figure 8: Percentile mean IRT scores in South Africa's 2016 PIRLS Literacy

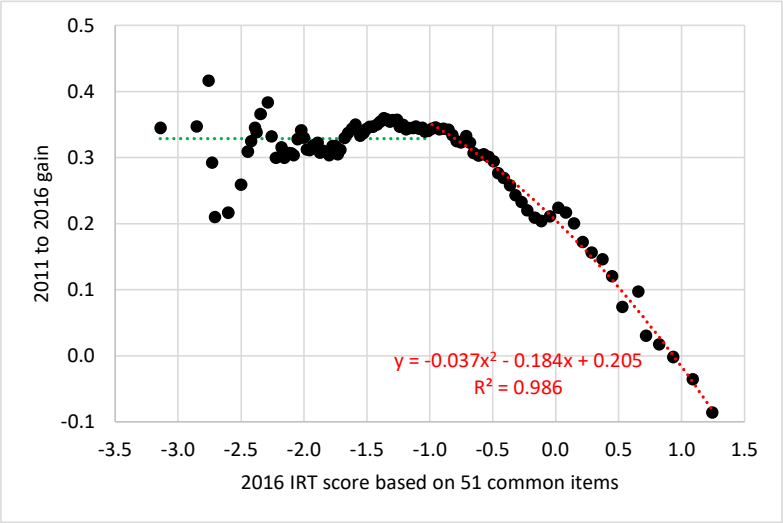


Note: Each marker represents a weighted percentile of pupils, sorted using the IRT score based on 51 common items.

For Figure 9 below, means per performance percentile in 2016 were compared to the corresponding percentile values for 2011, using the scale based on the 51 items throughout. The assumption used was of course that both samples were truly representative of the target population. Using this assumption, Figure 9 illustrates the gains experienced over the five years per percentile of the South African target population (there are 100 markers). Clearly, worse performing parts of the schooling system improved most. But gains hit a ceiling of

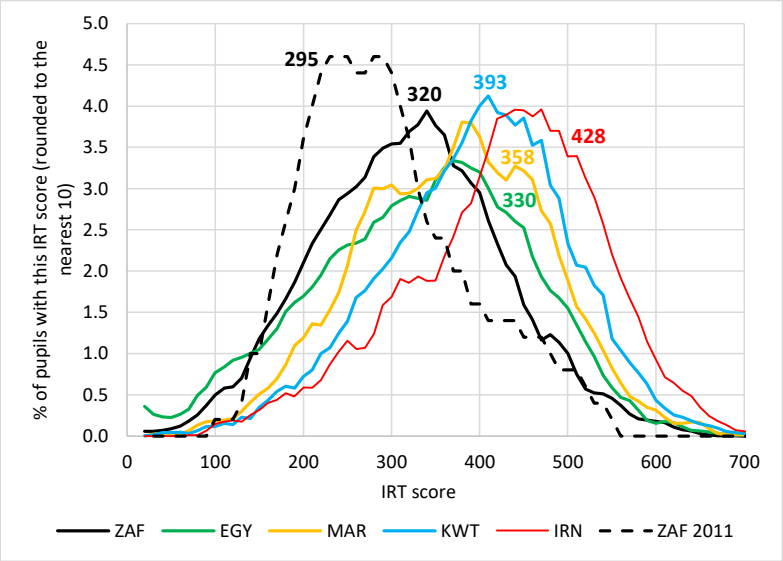
around 0.33 IRT points, or a third of a standard deviation. Outliers at the very bottom end of the performance distribution can be ignored.

Figure 9: 2011-2016 gains per South African percentile



An algorithm for converting the IRT scores based on 51 common items to the main PIRLS scale was developed using equations from both Figure 8 and Figure 9. First, a smoothed 51-item distribution across percentiles was produced for 2011, using the trendlines in Figure 9. For pupils scoring above -1.0 in 2016, the formula appearing in Figure 9 was used. For each of the 38 percentiles in question, the 2016 mean was taken, and the 2011 to 2016 gain was subtracted, using the Figure 9 equation, to obtain the percentile-specific performance in 2011. For pupils scoring below -1.0 in 2016, 0.329 IRT points were subtracted from the 2016 percentile-specific mean. This eliminated the effects of outliers to the left of Figure 9. Secondly, the 100 new IRT scores obtained for 2011, one per percentile, based on the 51-item scale, were transformed to the main PIRLS scale using the equation in Figure 8. This resulted in a mean of 294.8 across the 100 percentiles. The percentile-based distribution for 'ZAF 2011' is illustrated in Figure 10 below.

Figure 10: PIRLS Literacy IRT score distributions with ZAF 2011 included



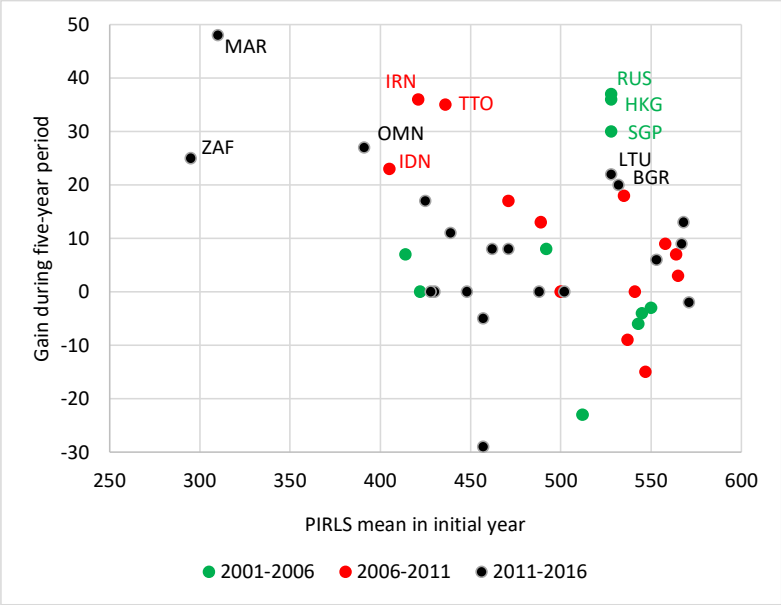
Note: The additional ZAF 2011 curve is smoothed using five, not three, values (two on either side of the original mean). This largely explains why on the far left, ZAF 2011 is below ZAF.

The difference between 295 in 2011 and 320 in 2016, of 25 PIRLS points, translates to an improvement of 0.05 standard deviations a year, using the 2016 standard deviation for South Africa of 106 standard deviations a year[9]. This is close to the 0.06 standard deviations a year based on the 51-item scale and discussed above.

How large is an improvement of 25 PIRLS points compared to other improvements reported in the PIRLS 2016 international report? This would be third-largest 2011 to 2016 improvement, after those of Morocco and Oman. Put differently, *South Africa's gain was the third-largest of the 43 gains for all countries with 2011 to 2016 gains reported in the PIRLS international report***.** If one also considers earlier improvements appearing in the 2016 international report, there were other noteworthy improvements, larger than the South Africa gain. For example, Iran and Trinidad and Tobago saw improvements of over 30 PIRLS points in the 2006 to 2011 period. Details appear in Figure 11 below. Clearly, *while South Africa's 2011 to 2016 gain was relatively large, even the 2016 mean of 320 remained low relative to those of other PIRLS participants***.**

[9] Mullis *et al*, 2017: 321.

Figure 11: PIRLS improvements 2001 to 2016



Note: Countries included are countries other than high-income OECD countries, in order to facilitate a comparison to other developing countries.

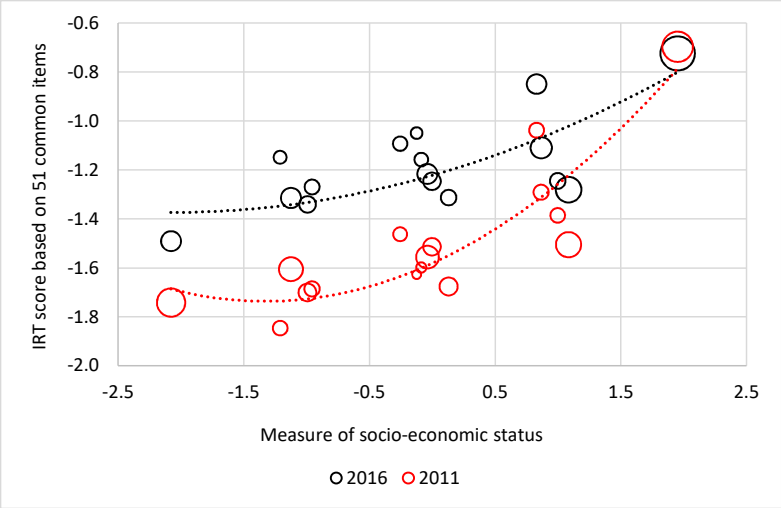
6 Taking into account the possibility of biases in the sampling

When PIRLS reports on country trends, it is assumed that in each year weighted pupils represent the same population, or that samples are sufficiently comparable over time. The analysis up to this point has used the same assumption. However, it is useful to interrogate this assumption a bit, and also to examine gains by socio-economic status. This additional analytical step can help to reassure that differences in means over time are not simply the result of a shifting sample, for instance a shift to a sample where the middle class is over-represented. Such a shift could create the illusion of a gain.

A measure of socio-economic status (SES) using the full set of 2011 data was first constructed. This was done using a principal components analysis and four binary variables describing access to household items: a computer; one's own study desk; one's own books; and one's own room. The resultant SES measure consisted of 16 different values, representing 16 combinations of the four variables. Comparable measures of SES were then produced for 2016, using the parameters obtained in the 2011 analysis. The measure returned a weighted mean of -0.018 in 2011 and 0.286 in 2016. In 2016, 89% of unweighted pupils had the required data to calculate the measure, against 94% in 2011. The change in the mean amounts to an increase of 0.25 standard deviations – standard deviations were 1.27 in 2011 and 1.19 in 2016. Such a change does not seem impossible. For instance, access to computers in the home is likely to have gone up.

However, to verify whether the PIRLS gain is driven by a different sample, or by actual qualitative gains, it is not necessary to examine the credibility of the SES change. A regression analysis that controls for SES will do the job. To introduce the analysis, Figure 12 illustrates the socio-economic gradients for the two years. Clearly, the largest gains were indeed seen for the poorest children, as suggested by Figure 9 above.

Figure 12: Percentile mean IRT scores in South Africa's 2016 PIRLS Literacy



Note: Curves are quadratic functions using weighted pupil observations. The number of unweighted pupils represented is 14,813 in 2011 and 11,372 in 2016.

For each of regression models A and B shown in Table 6, pupil weights were adjusted so that the total was equal across both years. Model A reflects a bivariate regression with no SES controls. This model points to a gain of 0.29 points along the scale using 51 items. The 95% confidence interval is 0.16 to 0.42, confirming that some gain is highly likely. Model B brings in the measure of SES. The gain drops a little to 0.26, but the confidence interval does not span zero. Even after controlling for socio-economic status, there is a statistically significant gain which is close to the gain seen if SES is not considered.

Table 6: Regression results with SES controls

	A	B
Dependent variable → IRT score based on 51 common items		
Constant	-1.477***	-1.545***
Is 2016	0.286*** (0.160-0.412)	0.262*** (0.148-0.377)
Assets		0.201***
Assets squared		0.066***
N	20,974	19,315
Number of schools	634	634
R^2	0.018	0.091

*Note: Values in brackets are 95% confidence intervals. *** indicates that the estimate is significant at the 1% level of significance. Levels of significance and confidence intervals take into account the design effects of clustering by school.*

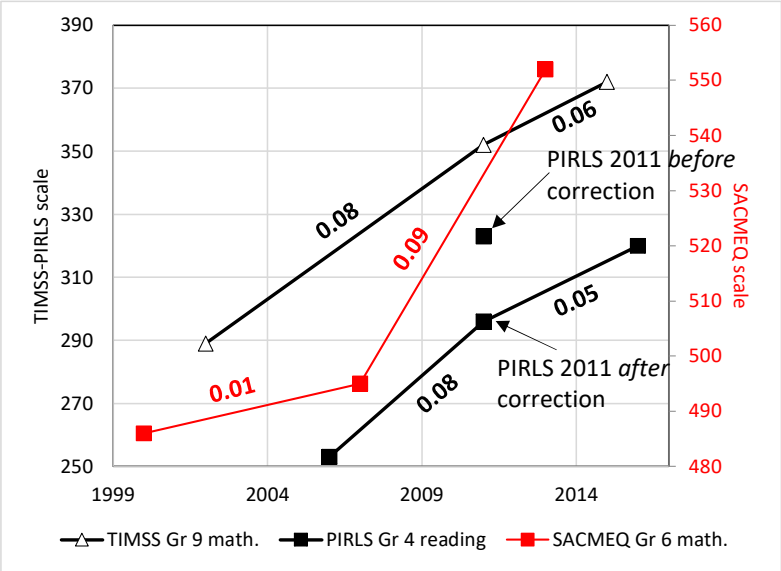
7 Conclusion

This paper has provided details on why the originally published 2011 to 2016 PIRLS trend for South Africa, of no progress, cannot be correct. In fact, *the analysis points to the gains being among the largest of all 2011 to 2016 gains, among 43 countries, reflected in the international PIRLS 2016 report*.

Figure 13 below indicates that the estimate for the correct 2011 South Africa mean, along the main PIRLS scale, of 295 points, produces a more plausible picture in the larger context. Firstly, it produces a 2006 to 2016 trend in PIRLS which is continuous, without an implausibly large gain between 2006 and 2011, followed by no gain between 2011 and 2016.

Moreover, a continuous gain in PIRLS would be in line with a continuous improvement in TIMSS across three points in time, and a large 2007 to 2013 gain in SACMEQ.

Figure 13: PIRLS, TIMSS and SACMEQ trends



Sources: The 2003 TIMSS Grade 9 mean of 289 is based on Reddy et al *(2016: 5) – the 2003 value of 285 in that source excluded private schools. TIMSS 2011 and 2016 values are from the international reports. SACMEQ values are from Makuwa (2010) and South Africa: Department of Basic Education (2017).*
Note: Values along the curves refer to annual standard deviation gains.

Detailed and policy-relevant analyses of assessment microdata are of course important. The question could be asked why analysts inside and outside government in South Africa (including the author of the current paper) did not interrogate the strangeness of the flat 2011 to 2016 earlier, when the trend was released. Answers would include insufficient capacity for this kind of work in South Africa, a sense that the 'black box' of assessment statistics are difficult to understand, and an insufficient realisation that educational quality *trends*, and not just cross-sectional analyses of the situation at one point in time, lie at the heart of a country's educational development.

References

All the reports in this list were freely available online in early 2020.

Gustafsson, M., Mabogoane, T & Taylor, N. (2012). *Where to from here: From fact to act.* Pretoria: Department of Basic Education.

Howie, S., Venter, E., Van Staden, S., Zimmerman, L. & Long, C. (2008). *PIRLS 2006 summary report: South African children's reading literacy achievement*. Pretoria: Centre for Evaluation and Assessment.

Howie, S., Combrinck, C., Roux, K., Tshele, M., K., Mokoena, G. & McLeod Palane, N. (2017). *PIRLS literacy 2017: Progress in International Reading Literacy Study 2016: South African Children's Reading Literacy Achievement*. Pretoria: Centre for Evaluation and Assessment.

Makuwa, D.K. (2010). Mixed results in achievement. *IIEP Newsletter,* XXVIII(3).

Martin, M.O., Mullis, I.V.S., Hooper, M. (eds.) (2017). *Methods and procedures in PIRLS 2016.* Chestnut Hill: IEA.

Mullis, I.V.S., Martin, M.O., Foy, P. & Hooper, M. (2017). *PIRLS 2016 international results in reading*. Chestnut Hill: Boston College.

Mullis, I.V.S., Martin, M.O., Foy, P. & Drucker, K.T. (2012). *PIRLS 2011 international results in reading*. Chestnut Hill: Boston College.

Mullis, I.V.S., Martin, M.O., Kennedy, A.M. & Foy, P. (2007). *PIRLS 2006 international report*. Chestnut Hill: Boston College.

Reddy, V., Visser, M., Winnaar, L. & Arends, F. (2016). *TIMSS 2015: Highlights of mathematics and science achievement of Grade 9 South African learners*. Pretoria: HSRC.

South Africa: Department of Basic Education (2017). *The SACMEQ IV project in South Africa: A study of the conditions of schooling and the quality of education*. Pretoria.

Taylor, S. & Taylor, N. (2013). *Learner performance in the National Schools Effectiveness Study*. Pretoria: Department of Basic Education. Available from: <https://www.researchgate.net> [Accessed January 2020].

UNESCO (2019). *How fast can levels of proficiency improve? Examining historical trends to inform SDG 4.1.1 scenarios*. Montreal.