



March 2023

Education is improving – but the message is hard to convey

Martin Gustafsson, Stellenbosch University

South Africa's improvements in international standardised tests have been exceptionally steep, albeit off a low base, for some 15 years. Six reasons make this difficult to communicate: examinations distract from reliable measurement; small gains are incorrectly considered insignificant; the social benefits of gains are long-term; the psychometrics behind rigorous testing is foreign to many; this testing has on occasion been poorly done; and teachers often express ideological concerns about standardised testing.

Introduction: what test trends tell us

A small revolution in how we think about schooling occurred around 2000. Economists, and then others, realised that in education what drove development was not so much levels of participation in school, but what children learn. While this may seem to state the obvious, before international test data became widely available it was not known how vastly the learning embodied in, say, six years of schooling varied across schooling systems.

South Africa is a global outlier for two reasons. Today, and even more so two decades ago, the basic reading and numeracy skills of South Africa's learners are strikingly low. This is now widely known. Less well known is that South Africa has been improving at a speed that is rapid by international standards. For instance, between 2011 and 2016 South Africa's progress in reading among primary learners was the second-fastest, after Morocco's, among countries in the Progress in International Reading Literacy Study (PIRLS).¹

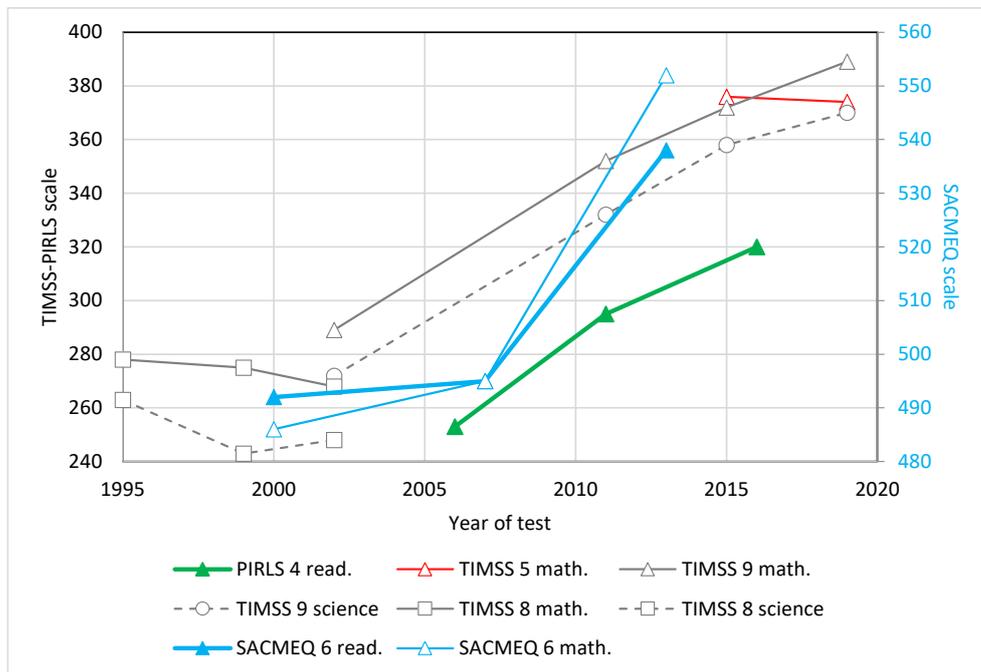
The following graph, from Gustafsson and Taylor (2022), illustrates the three international testing programmes all pointing to progress in South Africa. The TIMSS² Grade 5 mathematics trend 2015 to 2019 is a worrying anomaly as it suggests progress may have stalled at the primary level in the years immediately prior to the pandemic, even if this did not occur at the secondary level. But the overall picture is one of progress. The pandemic

¹ Van Staden and Gustafsson, 2023

² Trends in International Mathematics and Science Study.

was of course devastating for learning, and would have eliminated some of the past gains, though by no means all.

Results from international tests 1995-2019



Signs of progress are good for two reasons. Firstly, they are encouraging for the half a million people working in South Africa’s schooling sector. Secondly, they confirm that something must be working. This can help avoidance of unnecessary policy instability, something schooling systems are prone to. To put it crudely, if people believe there is no progress, it is easier to argue for yet another overhaul of the curriculum.

This article focusses on why this evidence of educational improvement is ignored by so many, which in turn leads to more pessimism than warranted, and a less informed debate about what it takes to improve in future. Mention of the available evidence is non-existent to scarce in the media and the policy documents of teacher unions. A recent report on South Africa’s education challenges, by the International Monetary Fund,³ an organisation one would expect to gravitate towards empirical evidence, barely discusses test-score improvements. While many in the 10 education departments, several academics, and a few journalists understand the improvements and their significance, reactions are often sceptical, even among government officials not working directly in the education sector.

Why? Six inter-related explanations are discussed below. While South Africa is the focus here, education analysts in other countries are likely to find them familiar.

³ Mlachila and Moeletsi, 2019

Examinations as a distraction

The first explanation is that examinations, such as South Africa's Grade 12 'Matric', often overshadow other indicators of progress. This often takes the form of a rather pointless quarrel over whether the authorities are cheating by manipulating the statistics, or participation in the examinations, to fabricate improvement for political gain. The argument is pointless because, as experienced education analysts will explain, examinations are poor indicators of educational progress. They are not designed for this purpose. Instead, they are designed to provide evidence of learning for individual students, with sufficient comparability between students within and across cohorts, to decide who, for instance, to admit into a university's engineering faculty. This is quite different from measuring whether the *average* competencies of students in mathematics are improving or not, or are better in one province than another.

An obvious problem is that different types of learners decide to take, say, mathematics at different points in time, and in different locations. These changes over time in the composition of the student group are known as selection effects.

More difficult to explain is that even the most experienced examiner is unable to set two examinations at two points in time of a sufficiently similar level of difficulty to gauge systemic progress. Moreover, no amount of statistical adjustment after the marking process can properly achieve this task. Yet a widely held expectation is that all this *is* possible. Unfortunately, the expectation is in part kept alive every time the education authorities speak about Grade 12 results as if they reliably reflect system-wide progress in the quality of learning. They can reflect progress with respect to the proportion of youths with information to navigate post-school studies and the labour market, but that is a different matter.

Inconvenient truths

If a first inconvenient truth in schooling is that traditional assessment techniques are inadequate to gauge qualitative progress, a second is that when system-wide progress occurs, it is painfully slow. If interventions focussing only on a selected group of schools typically struggle to shift the needle with respect to learning outcomes, it should come as no surprise that the Titanic that is an entire schooling system is difficult to steer towards better waters. Analysis by UNESCO (2019) suggests those developing countries that are educationally ambitious enough to participate in demanding global testing programmes, experience gains of around two percentage points a year in the proportion of young children displaying minimum competencies. If one views that against the fact that only around a quarter of South Africa's children display these competencies, getting all children up to speed is a 40-year project.

This slowness underlines the importance of truly rigorous measurement. If satisfactory progress means tiny steps forward year after year, then monitoring systems must be sufficiently fine-tuned to detect this.

These sobering realities clash with the very different world of targets, such as those of the United Nations Sustainable Development Goals. These targets point to 100% of children learning sufficiently by 2030. Unlike, for instance, targets around greenhouse gas emissions, these learning targets are purely aspirational and have no basis in empirical modelling. Wildly optimistic learning targets can do considerable harm to the careful evidence-based planning we need, and can make people dismissive of interventions that actually work fairly well.

The UN could never proclaim that, say, 60% of the world's children will be able to read properly by 2030. This would be politically unpalatable. Instead, the solution is to exalt small steps forward, and programmes that are able to measure this. We should think about it as follows: if the percentage of children displaying minimum reading skills rises from, say, 25% to 40% in 10 years, this should be seen as a success story, regardless of what the targets were.

The difficulty of the longer term

A third explanation is that the benefits of better learning in Grade 1 only really start to be felt in society some 15 years later, when, for instance, it becomes clear that younger electrical engineers are more productive than their older colleagues were because their educational foundations are stronger. This timespan exceeds the typical election cycle. This in part explains why generating excitement among politicians and voters around, say, improvements in the PIRLS programme is so difficult. Increasing the number of places for electrical engineering students at vocational colleges is easily a more viable political target.

A relevant link needs to be made between the role of foundational skills in making, for instance, the average electrical engineer more productive. cursory analysis of the distribution of foundational skills across countries makes it clear that the average electrical engineer in Germany or China would be better equipped to read manuals, understand algorithms, and specialise further than his or her peer in, say, South Africa, Paraguay or Bangladesh. A question that is at least as important as the number of youths entering vocational or technical education, is the degree to which foundational skills in earlier grades facilitate learning in these fields.

The challenge of psychometrics

Fourthly, methods in rigorous trend-focussed monitoring of learning are complex, and people often find it difficult to embrace what they do not fully understand. The psychometrics used in virtually all modern standardised testing systems would be foreign to most teachers. Yet they could be communicated and promoted better. Essentially there are two things that make standardised testing strange to many. Firstly, different students write different tests, consisting of different combinations of questions from a wider pool of questions. This, known as matrix sampling, has the benefit of testing across a wider range of skills, relative to having all students write the same test. Secondly, item response theory (IRT) is used to establish the difficulty of every question and thus to score students

consistently. Some questions, the so-called secure anchor items, are repeated across testing cycles, and must thus be kept secret. This security requirement is part of the reason why most standardised testing occurs in samples of schools, not the entire schooling system. Sampling must of course be strictly random for comparability to be sufficient.

Those wanting to understand the process, without become psychometricians themselves, have little to turn to.⁴ A series of manuals released by the World Bank from 2008 to 2015 aimed at policymakers inexplicably limits itself to how one conducts a standardised test *once*. They do not deal with the complexities relating to making tests results comparable over time. Guides that fill this gap are still to be written, it appears.

For those wanting to become psychometricians, opportunities for training in South Africa, and across most developing countries, are severely limited. South Africa, like most developing countries, probably does not have the local expertise needed to design testing programmes. This gives rise to an uneasy and risky over-reliance on experts from developed countries. In the larger scheme of things, training 10 or so South Africans to be excellent psychometricians should be a relatively easy task.

Fifthly, though the rise of national and international standardised testing in the past two decades has undoubtedly benefitted education planning, on occasion bad practices have led to the release of incorrect trend information, policy confusion, and opportunities for sceptics to feel vindicated. One of the best documented instances of this is not from a developing country, but from the United Kingdom, where what is now considered a non-existent decline in educational quality ignited an unnecessary blame game between Conservatives and Labour.⁵

South Africa has experienced more than its fair share of inaccurate measurement, but in both major instances where this occurred, corrections were made fairly timeously. In 2016, preliminary results from the SACMEQ⁶ programme were presented to a parliamentary committee. These results reflected an unbelievably large improvement between 2007 and 2013. Following concerns about the accuracy of this and about a lack of transparency around methods used, the data were revisited and new trends published. In the case of South Africa, the steepness of the improvement was halved. It was still impressive, but now also more believable within the wider historical context.

PIRLS initially released a flat 2011 to 2016 reading trend for South Africa, but revised this to an exceptionally steep trend after analysts interrogated the underlying data, which PIRLS makes easily available to researchers.⁷ The ideal is of course for these blunders not to occur, but a second-best option is for there to be a sufficient number of analysts within individual countries to raise the alarm when they do.

⁴ Gustafsson, 2019

⁵ See Jerrim, 2013

⁶ Southern and Eastern Africa Consortium for Monitoring Educational Quality.

⁷ Van Staden and Gustafsson (2023); Gustafsson (2020)

Ideological objections to standardised testing

Sixthly, standardised testing is sometimes said to carry undesirable ideological baggage. Teachers are a large and often politically powerful force with a sense of responsibility for reshaping society, sometimes radically. Education International, the world federation of teacher unions, has at times warned against international testing programmes, seeing them as part of a wider package of neoliberal reforms encompassing privatisation and the undermining of the employment tenure of teachers. But testing is not always placed within this 'neoliberal package'. Both socialist France and communist Cuba have introduced standardised testing.⁸ Critics easily see teacher opposition to testing as a push against accountability. But there are understandable concerns on the part of teachers. Standardised testing is generally focussed on a narrow set of basic skills in language and mathematics, which could lead to an under-valuing of less measurable skills, such as interpersonal skills and a sense of social responsibility. The counter-argument would be that these other skills can be properly realised only if the foundations are in place.

Opposition to standardised testing will often draw from well-documented and rather alarming experiences with testing in the United States, especially that conducted by Pearson Education, a major service provider. These experiences do indeed present a cautionary tale of the importance of government regulation and having the teaching profession involved in decisions. But standardised testing done well seems necessary. Without it, a schooling system is like a ship without navigation equipment.

Despite the strong presence of broader ideological arguments in the policies of many teacher unions, education authorities tend to interact with unions as if these broader concerns did not exist. Perhaps a more productive interaction would be possible if these concerns were put on the agenda, especially in countries such as South Africa where the general political rhetoric of government and unions is fairly similar. For instance, when is testing a threat, and when is it not? Which arguments apply to sample-based testing, and which to censal testing (testing that covers every learner in a specific grade)? What are the risks to disadvantaged communities where there is an absence of censal testing?

Conclusion

If we are serious about improving the quality of education, we need to understand how this is measured, and pay careful attention to relevant test statistics. While debating Grade 12 examination results is a firmly entrenched habit in South Africa, this is not where we should be looking for the answers. Rather, existing sample-based testing systems have been providing answers for the past two decades. Of great importance is the government's new sample-based Systemic Evaluation programme, the results of which are expected in 2023. This could create the baseline for a new time series of comparable statistics. What still seems missing, however, is a clear commitment to develop a critical mass of South

⁸ Gustafsson, 2019

African psychometricians, so that the current over-reliance on foreign experts can be addressed.

References

Gustafsson, M. (2019). *The case for statecraft in education: The NDP, a recent book on governance, and the New Public Management inheritance*. Stellenbosch: University of Stellenbosch.

Gustafsson, M. (2020). *A revised PIRLS 2011 to 2016 trend for South Africa and the importance of analysing the underlying microdata*. Stellenbosch: Stellenbosch University.
Gustafsson, M. & Taylor, N. (2022). *The politics of improving learning outcomes in South Africa*. Oxford: RISE.

Jerrim, J. (2013). The reliability of trends over time in international education test scores: Is the performance of England's secondary school pupils really in relative decline? *Journal of Social Policy*, 42(2): 259-279.

Mlachila, M. & Moeletsi, T. (2019). *Struggling to make the grade: A review of the causes and consequences of the weak outcomes of South Africa's education system*. Washington: IMF.

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

Van Staden, S. & Gustafsson, M. (2023). What a decade of PIRLS results reveals about early grade reading in South Africa: 2006, 2011, 2016. In Nic Spaull and Elizabeth Pretorius (eds.), *Early grade reading in South Africa*. Cape Town: Oxford University Press.