Economics of Education Review 62 (2018) 267–286

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

Internationally comparable mathematics scores for fourteen african countries

Justin Sandefur[1]

Center for Global Development, Washington, DC 20002, United States

ARTICLE INFO

Keywords:
Learning assessments
Education quality
Human capital
Africa

JEL classification:
I25
J24
O15
O55

ABSTRACT

Internationally comparable test scores play a central role in both research and policy debates on education. However, the main international testing regimes, such as PISA, TIMSS, or PIRLS, include almost no low-income countries. Instead, many developing countries opt for regional assessments sponsored by the United Nations. This paper exploits an overlap between the regional test for Southern and Eastern Africa, SACMEQ, and the TIMSS test – in both country coverage, and questions asked – to assesses the feasibility of constructing global learning metrics by equating regional and international scales. I find considerable variance when comparing three commonly-used equating methods, suggesting precise country rankings are unreliable. Across all methodologies, however, learning levels in this sample of African countries are consistently (a) low in absolute terms, by roughly one-and-a-half standard deviations or more compared to OECD pupils of a similar age; (b) significantly lower than predicted by African per capita GDP levels; and (c) converging slowly, if at all, to the rest of the world during the 2000s. The robustness of these simple facts suggests even crude linking methods may suffice for many international policy questions, such as tracking the UN's development goals.

1. Introduction

Around the developing world, and particularly in East Africa, there is growing evidence that the expansion of primary school enrollment over the last quarter century has not delivered concomitant improvements in learning levels (Jones, Schipper, Ruto, & Rajani, 2014; Pritchett, 2013). The United Nation's 2015 global goals seek to address this imbalance by focusing on education quality, including indicators of proficiency in literacy and numeracy. But these indicators are not currently measurable on an international scale, particularly in Africa. Notably, only three countries in Sub-Saharan Africa have participated in any of the major international assessments of learning levels.[2]

My goal in this paper is to put mathematics test scores from an existing regional learning assessment covering fourteen African countries on an international scale using both simple statistical methods, and more formal item response theory methods. This process is known in the psychometric literature as linking or equating, terms which I use interchangeably here.[3] The regional test is the Southern and Eastern Africa Consortium for Measuring Education Quality (SACMEQ) assessment, and the international scale is provided by the Trends in International Mathematics and Science Study (TIMSS), an international assessment administered in grades three and four (population 1) and seven and eight (population 2) in over sixty countries. This linking is possible because (a) two countries, Botswana and South Africa, took both tests, and (b) the 2000 and 2007 SACMEQ rounds embedded a number of items from the TIMSS test. These overlapping items were included in the African tests with the explicit purpose of facilitating international comparisons (Ross et al., 2005, p. 71).

It appears this *ex ante* push for comparability was abandoned *ex post*. To my knowledge, no reporting of SACMEQ scores on an international scale exists in the public domain. It is widely rumored that these results were withdrawn due to the political sensitivity of highlighting the enormous learning deficiencies in all fourteen SACMEQ countries relative to the global distribution. There is mixed evidence to justify this political sensitivity. There are anecdotal reports that benchmarking student performance on international assessments has

E-mail address: jsandefur@cgdev.org.

[1] The analysis here was conceived in discussions with Luis Crouch, Beth King, and Nic Spaull, and has benefited enormously from comments on an earlier draft from Barbara Bruns, Luis Crouch, Jishnu Das, Deon Filmer, Lant Pritchett, and (in particular) Abhijeet Singh. Maryam Akmal and Dev Patel provided excellent research assistance. The views expressed here do not necessarily reflect those of the Center for Global Development, its board, or its funders. All errors are mine.

[2] Ghana has participated in PIRLS (primary-level reading assessment), and South Africa and Botswana have participated in both PIRLS and TIMSS (primary-level mathematics and science assessment).

[3] See Holland (2007) for a discussion of what makes a linking an equating; the latter generally implies greater rigor and comparability. In Holland's terminology, this exercise might be termed a 'calibration' or 'concordance', as the pupil populations differ and the test constructs, difficulty, and reliability are not guaranteed to be identical.

https://doi.org/10.1016/j.econedurev.2017.12.003
Received 4 January 2017; Received in revised form 30 October 2017; Accepted 17 December 2017
Available online 26 December 2017
0272-7757/

contributed to national political pressure for education reform in the OECD (Breakspear, 2012), as well as some Latin American (Bruns, 2015) and Eastern European countries (Marciniak, 2016). But experimental work in East Africa has found that dissemination of national assessments results has little effect on local political demands for education reform (Lieberman, Posner, & Tsai, 2014).

Politics aside, there are sound technical reasons to be cautious about any comparison of African learning levels to international benchmarks. When comparing populations with very different learning levels, traditional methods for test-score equating are subject to sizable non-sampling error. The size of this 'linking error' is inversely proportional to the number of overlapping items across the two the tests (Michaelides & Haertel, 2004). For instance, Hastedt and Desa (2015) present simulations using TIMSS data to show that statistically significant differences in country means may not be detected when the number of overlapping items falls below roughly thirty, as is the case here. As noted below, however, the magnitude of statistically significant discrepancies may be small when compared to the true learning gaps between, say, many sub-Saharan African countries and an OECD sample.

To address these concerns, I compare the results from three different linking approaches.

The first approach is referred to as equipercentile equating or linking in the psychometric literature (see Kolen and Brennan, 2014, chapter 4). It does not require any overlapping test items across the two tests and does not rely on item response theory to link the two test scales, above and beyond whatever IRT methods may have been used in construction of the original scores. Instead, equipercentile linking as applied here depends on the existence of data from both tests for a common population of pupils. In this case, I rely on overlapping coverage of SACMEQ (2000) and TIMSS (2003) in Botswana and South Africa, matching each percentile of the SACMEQ distribution to the corresponding percentile of the TIMSS distribution. Lee and Barro (2001), Altinok and Murseli (2007), and Altinok, Diebolt, and Demeulemeester (2014) have all applied simpler versions of this approach to link various regional and international tests, relying only on country means and variances; here I apply non-parametric methods to the full distribution and take a more conservative approach to identifying comparable populations of test-takers. Nevertheless, this procedure assumes that SACMEQ and TIMSS true scores are highly predictive of each other, and that this relationship is stable across countries. The first assumption is not testable with my data, and I find some violation of the second assumption when comparing results for Botswana and South Africa.[4]

A second, alternative approach using item-response theory relies on overlapping items across the two tests, rather than overlapping coverage in the populations tested. Das and Zajonc (2010) apply IRT methods to estimate TIMSS-equivalent scores for two states in India, and Singh (2014) applies the same procedure to regions of Ethiopia, India, Peru, and Vietnam. The two central assumptions here, as in most applications of item response theory, are unidimensionality of the underlying trait (which I refer to as math proficiency) and parameter invariance (e.g., that the relative difficulty of different items is stable across populations). The linking procedure implicitly assumes SACMEQ and TIMSS measure not only a unidimensional trait, but that it is the same trait. Violations of these assumptions manifest themselves through differential item functioning (DIF), in which students with similar proficiency levels in different groups (in this case, the SACMEQ African sample versus the broader TIMSS sample) perform better or worse on a given item. While teachers in the SACMEQ sample pool quite well with the TIMSS sample, SACMEQ pupils exhibit high levels of DIF, casting some doubt on these estimates, which are considerably higher than the other two approaches – and well above the actual TIMSS scores measured for Botswana and South Africa.

A complication to this approach is that the SACMEQ pupil test includes only a few TIMSS items; however, the SACMEQ teacher test includes a longer list of TIMSS items, and the SACMEQ teacher and pupil tests also share a longer list of items between them. (Jump ahead to Figure f:venn8 for an illustration of the overlap.) Thus I present an extension of standard linking methodologies, effectively creating a chain linkage from TIMSS to the SACMEQ teacher test and then, in turn, to the SACMEQ pupil test.

The third approach I employ also relies on item response theory, but is potentially less sensitive to DIF. This approach, known as mean-sigma equating, is commonly applied to link, e.g., subsequent rounds of testing regime. Rather than imposing all of the item level parameters from the reference population (TIMSS) on the target population (SACMEQ), it ensures only that the average level of difficulty and discrimination for the overlapping items are held constant across the two populations. Estimates based on the mean-sigma approach are largely congruent with the equipercentile method as well as the actual TIMSS scores for Botswana and South Africa.

Substantively, the results here are daunting for African education systems. When comparing SACMEQ pupils to TIMSS pupils of a similar age (roughly fourteen-years-old in both cases), most of the national test-score averages I estimate for the fourteen African countries in my sample fall more than two standard deviations below the TIMSS average, which places them below the 5th percentile in most European, North American, and East Asian countries. In contrast, scores from the SACMEQ test administered to math teachers are much higher, but fall only modestly above the TIMSS sample average for seventh- and eighth-grade pupils, in line with earlier analysis by Spaull and van der Berg (2013). The African scores for children of similar ages also appear low relative to national GDP levels; in a regression of average scores on per capita GDP in PPP terms, average scores in the SACMEQ sample are significantly below the predicted value using all three linking methodologies. Furthermore, there is little sign that African scores were improving rapidly or converging to OECD levels during the 2000s.

A major caveat in interpreting these comparative results is that the SACMEQ test is administered to pupils in grade 6 in most countries, while TIMSS is administered to pupils in grades 7 or 8. Thus African pupils are one to two grades below their OECD counterparts when sitting these tests. There is some virtue to this difference. Because pupils in the African sample tend to be much older at a given grade level, the average *age* of the pupils in the SACMEQ and TIMSS data is quite similar. In most countries, the modal age in each case is fourteen.[5]

One extremely conservative approach to acknowledging the grade difference between SACMEQ and TIMSS is to compare grade 6 pupils in SACMEQ to pupils who sat the grade 3 or 4 test for TIMSS, who are typically about four years younger than their African counterparts. By this measure, the top-scoring African countries produce grade 6 scores that are roughly equivalent to grade 3 or 4 scores in some OECD countries. For instance, Kenyan sixth-graders score just above New Zealand fourth-graders. But the bulk of the African sample still falls well short of their younger OECD peers, with African countries occupying eleven of the bottom fourteen spots on a combined SACMEQ–TIMSS league table using my linking results.

Methodologically, this exercise aims to clarify what can and cannot be reliably stated when linking regional and international learning scales, particularly when the link relies on a very short set of anchoring items and attempting to span populations with widely disparate

[4] An alternative approach to linking international assessments that has been used and widely cited in the economics of education literature (Hanushek & Kimko, 2000; Hanushek & Woessmann, 2012) abstracts entirely from the content of the test or the distribution of pupil scores, and uses assumptions about the variance of country averages around the world to link the global distributions of various assessments. See Altinok et al. (2014) for a critique of this approach.

[5] Note that while both TIMSS and SACMEQ use grade-based sampling, other major international assessments such as PISA use age-based sampling.