Best practice in the design of national assessments

A few myths and trade-offs

Martin Gustafsson (mgustafsson@sun.ac.za)

Department of Economics

Stellenbosch University

Version: 9 April 2013

Contents

- Introduction
- Myth 1: Broad-based governance is always necessary (or the question of what explains Brazil's success).
- Myth 2 (an easy one to overturn): Samples are about percentages.
- Myth 3: You can achieve comparability in results simply through test equivalence.
- Myth 4: Standardised testing with accountability pressures has led to untenable levels of cheating.

- Myth 5: Sudden and large improvements in performance are possible.
- Myths 6a and 6b: Information-only approaches can easily work / You must always couple assessment to support.
- What does all this mean for the design of ANA?

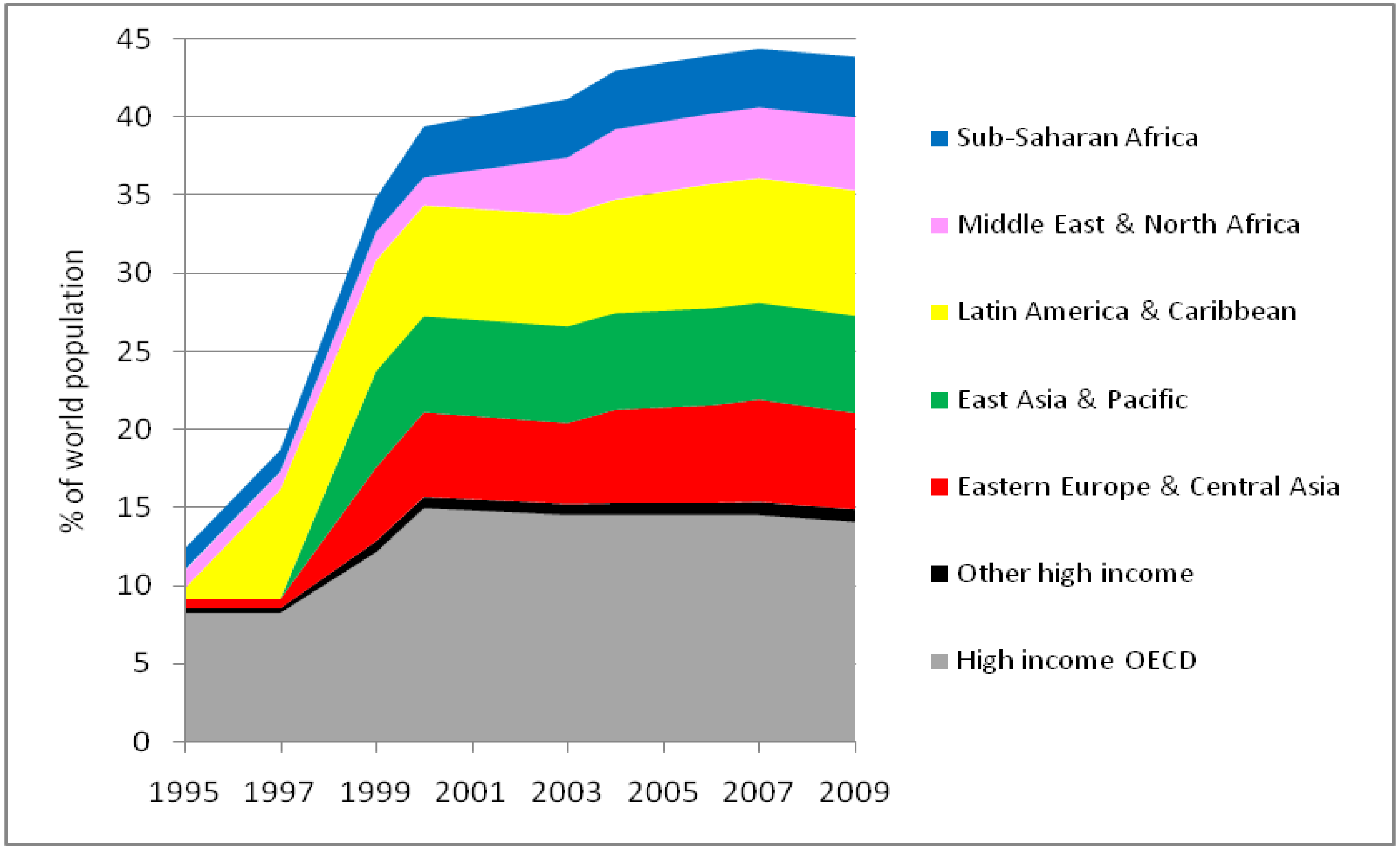
Note: This presentation is stand-alone, it does not come with a full narrative. For this reason points are made especially clearly (one hopes) and there are several references on the various slides to source documents for the student wishing to find out more.

Introduction

- National assessments have become an industry and a bit of a science. See for instance the World Bank's recent entrance into this somewhat non-economic arena. See the books at http://go.worldbank.org/M2O1YDQO90.
- There has been a noteworthy disjuncture between the debates on national assessments and debates on more traditional examinations. It's a pity that e.g. the role of the latter in promoting accountability and the often tenuous nature of the distinction are not better covered in the literature. The problems of using examinations as, in a sense, standardised assessments in South Africa is a focus of e.g. Taylor (2009) and Reddy (ed., 2006).

UNIVERSITEIT
STELLENBOSCH
UNIVERSITY

Which countries have entered some international standardised assessment programme



Graph produced by myself using relevant data.

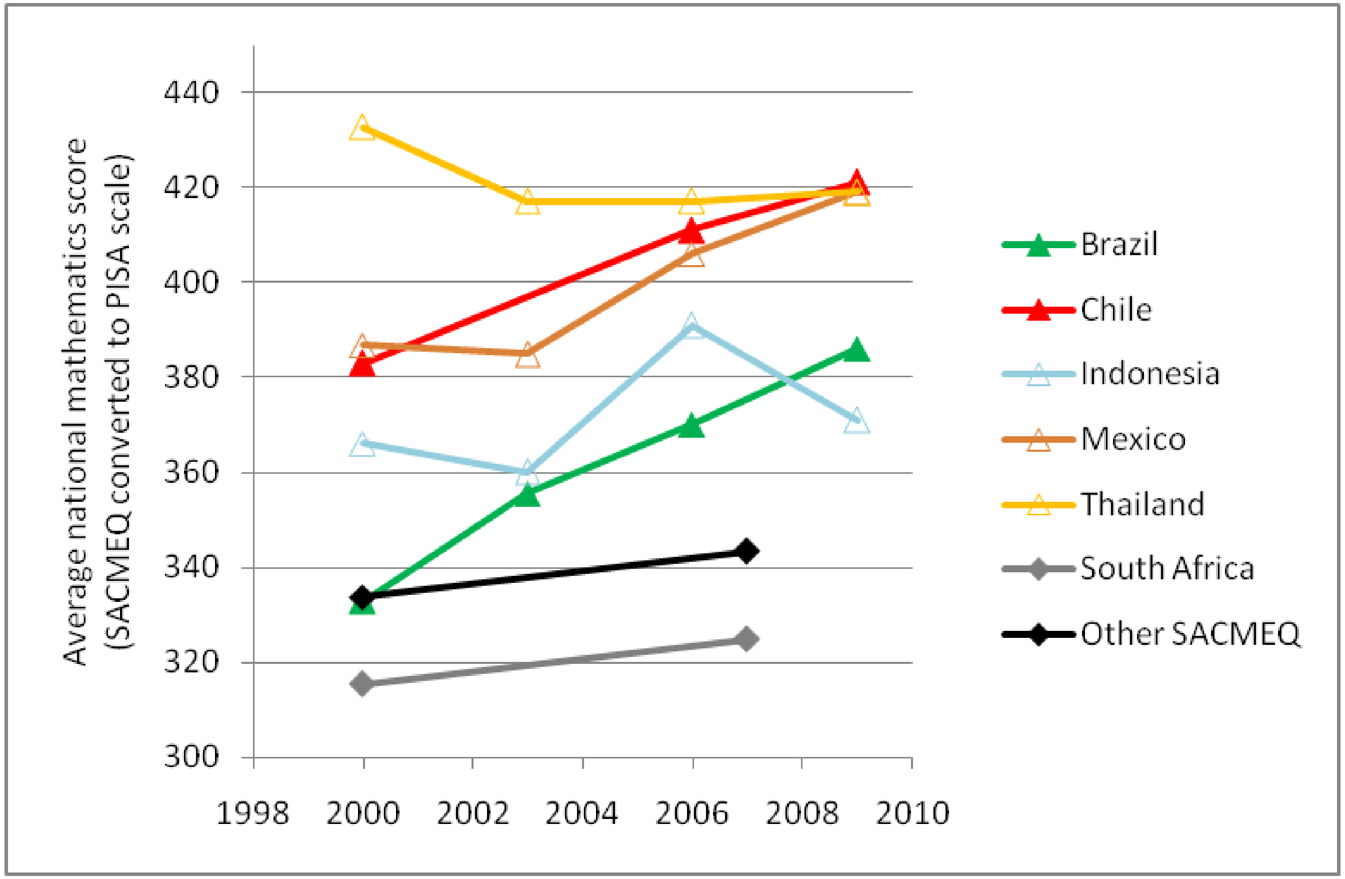
Most of the remaining 55% is India and China

Myth 1: Broad-based governance is always necessary (or the question of what explains Brazil's success)

- The World Bank guides are rather strong on multi-stakeholder participation in the governance of a national assessment (Greaney and Kellaghan, 2008).
- An excellent account of how a multi-stakeholder approach involving unions can strengthen the process is provided by Ravela (2005, 2006), in relation to Uruguay.

- However, Brazil proceeded rather differently. High levels of technical capacity at the national level have permitted a rather centralised implementation approach that is widely respected.
  - In a nutshell, the sample-based SAEB, a bit like our Systemic Evaluation, was introduced in 1990, and expanded to the two-tier sample plus universal Prova Brasil in 2005, rather like our ANA. I produced a report, titled *Quality enhancement options for the schooling system* in 2009, as part of a UNICEF-funded school funding study, where I put together key information about a number of testing systems and the implications for South Africa. The report is unfortunately not on the web, but e-mail me and I’ll send it.

Why Brazil is so important to watch



UNIVERSITEIT
STELLENBOSCH
UNIVERSITY

Myth 2: Samples are about percentages

- Even if one’s national assessment is universal, it is considered necessary to have a verification sample, with more stringent administrative controls than the universal component, but also with item-level capturing and background questionnaires.
- But how large does one’s sample need to be?
- A common misperception is that it is all about a percentage of the population, so a schooling system that is twice as large as another requires twice as large a sample.

- It may seem strange, but it is actually about an absolute number of sampled units. The size of the population plays very little role (though it does play a small role). This is why e.g. Botswana and the USA have almost the same number of sampled TIMSS pupils.
  - Perhaps think of it as follows: If you have a pot of soup, to find out what it tastes like you need to try it in a teaspoon. Using a tablespoon makes no difference. If the pot is larger, you don’t need a larger spoon.
- But how large should the sample be, in absolute numbers, then?

- What is widely used is the IEA's standard that results in e.g. 392 schools and around 9,000 learners in SACMEQ 2007. The results depend on how small you want your confidence intervals to be and the variation (inequality) in your data. The IEA standard is a bit difficult to find clearly stated (at least I had problems). It is implied in Ross (2005: 22) and more explicit in the document *Sample design procedures for the SACMEQ II project*.
- If you want to meet the IEA standard for provincial statistics, you need about 392 schools per province! This is virtually never achieved, so we live with large confidence intervals at that level.

UNIVERSITEIT
STELLENBOSCH
UNIVERSITY

- To what extent does the population size influence sample size? A power analysis will show that e.g. 125 schools in KwaZulu-Natal yields the same confidence intervals as 110 schools in Northern Cape. So a ratio of around 1.15, though population in KN is 9 timed as large as that in NC.

Myth 3: You can achieve comparability in results simply through test equivalence

- It’s a common belief: If you just get good enough test designers, you can design two different tests that will yield comparable results in a standardised assessment, using a simple marking approach. In ANA 2011 to 2012 this belief is implicit.
- Unfortunately, no team of test designers is this good!

- Testing systems tend to transcend this problem through two stages in their development:
  - First, you use an IRT (item response theory) approach in the marking process, though you still have two whole tests which you make as comparable as possible and which include anchor items. This we see in SACMEQ 2000 and 2007 (actually, the tests used in the two years were the same tests, so highly comparable!).
  - Second, you let pupils write different but more or less equally difficult versions of the test in each test run, with some common items (questions). This we see in e.g. the more recent runs of TIMSS. SACMEQ will apparently move in this direction in its next run.

- So how does IRT marking work?
  - You use the anchor items to see which pupils are at similar levels of achievement.
  - Then you grade the difficulty of non-anchor items on the basis of the anchor items. This occurs through a complex statistical approach, e.g. Rasch.
  - What the above means is that you use actual performance of pupils to adjust your assumptions around how difficult items are.

- It also means that you cannot have e.g. a simple mark of 54 out of 100. Instead you have the typical ‘mean is 500, standard deviation is 100’ approach of e.g. SACMEQ.

- And what about different versions of the same test in the same run?
  - The problem with the previous solution is that you need to have few anchor items to avoid problems associated with re-using virtually the same test (cheating!), but at the same time you need more anchor items to improve comparability.
  - The way out is to have several versions of the test within one run and let anchor items join versions within one run as well as different runs. Apart from tightening comparability through more anchor items, you also broaden the topics that you can cover.

- To illustrate, let’s see TIMSS 2003 mathematics: There were 194 test items, but each pupil took only around 40 items. There were 12 versions of the test, each with a different combination of the 194 test items.

Myth 4: Standardised testing with accountability pressures has led to untenable levels of cheating

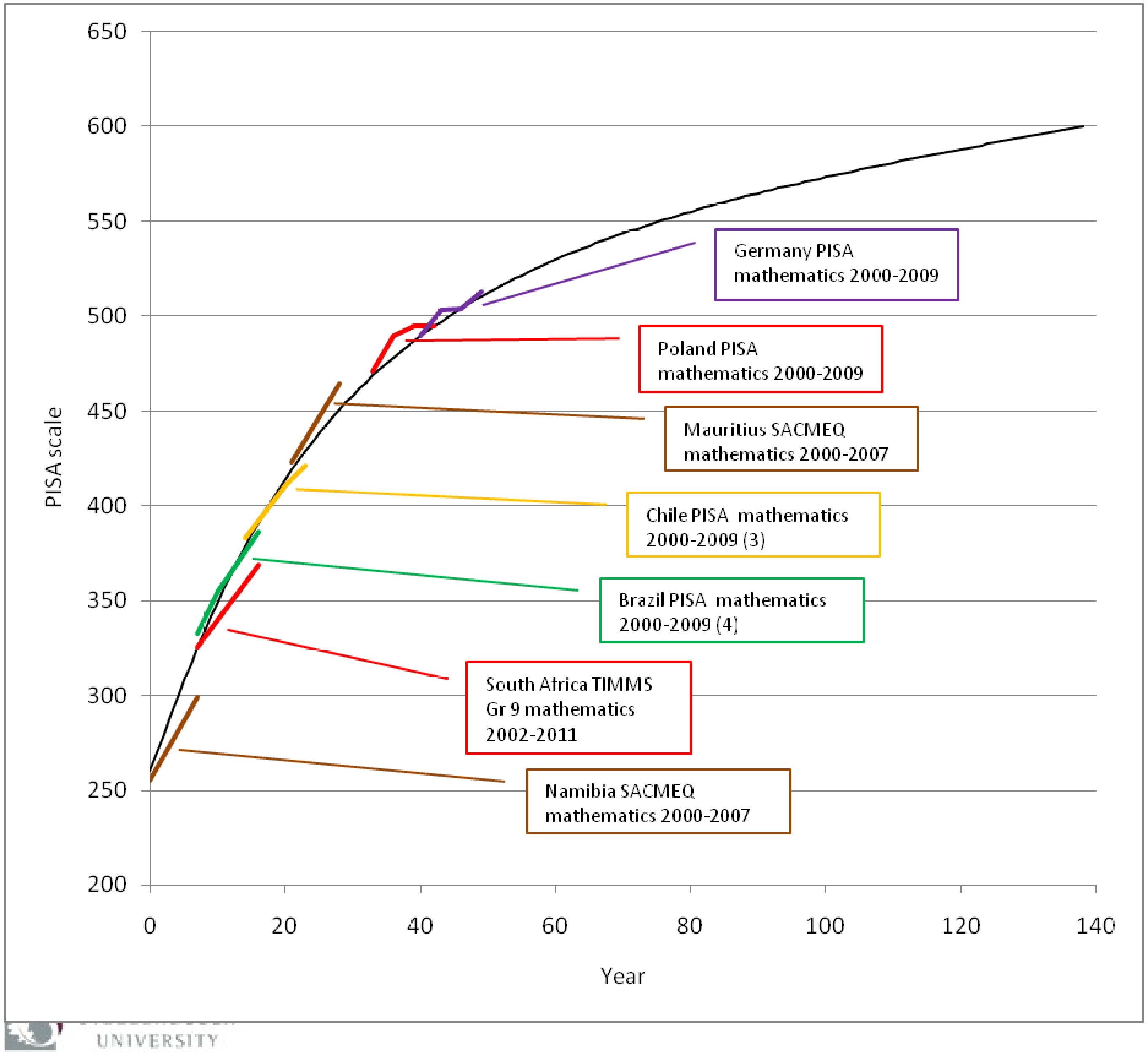
- Highly publicised reports (plus the movie *Freakonomics*) from the US have fed the notion that standardised testing leads to massive cheating which undermines the whole test programme.
- Clearly there is a problem, but we should not lose sight of its magnitude and whether experiences outside the US are different.
- Within the US, despite a few scandals, more systematic research does not point to widespread cheating undermining the process. See for instance Jennings and Rentner (2006).

- In developing countries, there are fewer reports of systematic cheating. A key factor could be that in these countries test administrators are often external to the school, something many First World teachers would find unacceptable. Brazil, for instance, has external administrators even for the universal tests written by all schools.

Myth 5: Sudden and large improvements in performance are possible

- It would be good if this were true, but...

The best possible improvement trajectories



This graph illustrates recent strong positive trends displayed by key countries with respect to standardised test results. The black line represents a best possible trend explained in Gustafsson (2012). The method for converting TIMSS and SACMEQ values to a PISA scale is explained in the same paper.

- It is useful to think of the best possible annual improvements in terms of standard deviations.
- The best possible is around 0.08 standard deviations up per year.
  - ➢ That’s about 8 SACMEQ points in a year.
  - ➢ If we compare a few ANA averages from 2011 and 2012, we see that both implied upward and downward ‘trends’ couldn’t be real. They are too large.

	Published average % score		Implied	Std. dev. in	Largest shift possible using
	2011	2012	shift	2011	0.08 s.d. criterion
Gr 3 math	**28**	**41**	**+13**	**20.0**	**+1.6**
Gr 6 math	**30**	**27**	**-3**	**17.3**	**-1.4**

Key texts

Greaney, V. & Kellaghan, T. (2008). *Assessing national achievement levels in education*. Washington: World Bank.

Gustafsson, M. (2012). *More countries, similar results: A nonlinear programming approach to normalising test scores needed for growth regressions*. Stellenbosch: Stellenbosch University. Available from: <http://ideas.repec.org/p/sza/wpaper/wpapers164.html> [Accessed August 2012].

Jennings, J. & Rentner, D.S. (2006). Ten big effects of the No Child Left Behind Act on public schools. *Phi Delta Kappan*, 88(2): 110-113.

Mullis, I.V.S., Martin, M.O., Foy, P. & Arora, A. (2012*). TIMSS 2011 international results in mathematics*. Chestnut Hill: Boston College. Available from: <http://timss.bc.edu> [Accessed December 2012].

Ravela, P. (2005). *A formative approach to national assessments: The case of Uruguay*. Prospects, 35(1): 21-43.

Ravela, P. (2006). *Using national assessments to improve teaching... and learning: The experience of UMRE in Uruguay*. Washington: World Bank.

Available from: <http://info.worldbank.org/etools/docs/library/235974/D2_cartagena_2006_PRavela_fv_11sep06.pdf> [Accessed November 2009].

Reddy, V., ed. (2006). *Marking matric: Colloquium proceedings*. Pretoria: HSRC. Available from: <http://www.hsrcpublishers.ac.za> [Accessed June 2007].

Ross, K.N. (2005). *Sample design for educational survey research*. Paris: IIEP. Available from: <http://www.iiep.unesco.org/> [Accessed April 2012].

Taylor, N. (2009). Standards-based accountability in South Africa. *School effectiveness and school improvement*, 20(3): 341-356.