

Caution! Using high stakes testing of student learning in development

Author : Robert Cannon

Date : October 25, 2012



This is the first of a two part series on high stakes testing. The second part can be found [here](#).

The Center for Global Development published a [Policy Paper](#) in September by Amanda Beatty and Lant Pritchett titled *From Schooling Goals to Learning Goals; How Fast Can Student Learning Improve?*

This is a welcome contribution to the discussion about what might replace the Millennium Development Goals, post 2015. The Paper is welcome because it foregrounds learning, moving away from *schooling* goals, such as enrolment and completion, and towards *learning* goals.

Beatty and Pritchett show that the pace of progress for developing countries in international assessments of student learning is very slow. They demonstrate how it would take 'forever' for some developing countries to reach current OECD levels. The data they use in their analysis are from the assessment of fourth and eighth graders in the *Trends in International Mathematics and Science Study – TIMSS*, the *Programme for International Student Assessment – PISA* – a study of 15-year-old's performance on mathematics, science, and reading, and [SACMEQ](#), *The Southern and Eastern Africa Consortium for Monitoring Educational Quality* for mathematics and reading.

Some might see the use of these tests as the perfect policy mechanism in development because they are thought to 'effect' or motivate change in the system, and, on the other, are supposed to 'detect' whether changes in the system have actually occurred. Seen in this way, could any way be better for development policy makers' intent on improving the quality of education? The short answer is 'yes'. One better way may be to avoid such tests for these purposes!

This answer demands an explanation. Beatty and Pritchett skip lightly over the nature of the indicators of learning they have used in their analysis and ignore the negative consequences of their use in development work. In my view, this is where we must proceed with great caution. To be fair, the authors insist that they are not wedded to the metrics of these tests and they encourage the use of country-specific goals and tests.

Nevertheless, they go on to 'encourage' countries to participate in these tests in order to facilitate comprehensive international comparisons (p.19). Here, the need for indicators threatens to trump the consequences – that can be very serious!

The authors discuss two of their concerns with using TIMSS, PISA, SACMEQ, or similar tests for international comparisons. First, these tests do not address age-group cohorts of children and are based only on sampling data of enrolled children. They argue that cohort-based learning goals should apply to all children, in or out of school. This seems sensible but will be very difficult to achieve in practice. Second, the tests only assess children in later grades whereas there are many stages from commencement to completion of compulsory education and many goals in these different stages.

These are reasonable technical concerns, but addressing them will have serious human consequences. Both concerns imply *more* testing of *more* children *more* often. Anyone who seriously believes this is a good educational idea should read either or both of: *The Paradoxes of High Stakes Testing: How they affect students, their parents, teachers, principals, schools and society*, by G. Madaus, M. Russell and J. Higgins, Charlotte, NC, Information Age Publishing, 2009 (available [here](#)) and *Collateral Damage: How high stakes testing corrupts America's schools*, by S.L Nichols and D.C. Berliner, Cambridge, Harvard Education Press, 2007 ([here](#)). These books are not popular polemics but serious, research based texts written by some of the most respected scholars in the field of educational testing.

Madaus and his colleagues conclude in Chapter 8 that high stakes testing produces “chronic unintended negative consequences” – narrowing the curriculum, decreasing attention on non-tested subjects, corrupting test results, cheating, retaining students in grade, increased drop out rates, among others. These authors consider ways in which technology might be applied to advance testing and argue for rigorous, independent monitoring of high stakes testing to ensure that benefits outweigh the harms. Nichols and Berliner go much further, concluding that they have no doubt these tests corrupt educators and harm schools. They express the hope that others will join them in demanding a moratorium on high stakes testing (p.202). Together, these books eviscerate the proposition that schools and systems can be fairly evaluated with current approaches. They provide damning evidence of the negative consequences for children, teachers and schools of current policies and the current state of knowledge that support high stakes testing.

I have two more concerns about these kinds of tests: the comparability of the indicators between countries, and the negative consequences that may flow from their use by developing countries.

The technical [research literature](#) on testing is raising doubts about the assumption that test performance is independent of the language of questioning and that linguistic, geographic, and socio-economic and cultural factors have an impact that makes meaningful international comparisons between countries problematic. Further, as the relationship between test performance, home, and social class factors in the USA alone calls into question the use of test scores to judge school quality ([here](#), page 62), then how valid can these kinds of tests be for comparative purposes between diverse countries and cultures to assess development progress?

My second concern is with the consequences for the whole educational enterprise of using these kinds of tests. For countries, these are 'high stakes' tests. Doing well on them and maintaining a consistent standard of performance over time is a matter of national prestige that has major political consequence – the Australian government's policy with [NAPLAN](#) (the National Assessment Program Literacy and Numeracy) and the [My School](#) web site is just one example.

Donald Campbell, an eminent American social scientist noted for his work in quantitative methodology, proposed in the 1970s what is now known as “Campbell's law”. This law has two parts, one of which is concerned with the validity of the test indicators we use, and another that is concerned with the effects on organisations and people that work with indicators of high stakes value. Campbell's law, as discussed [here](#) on page 4 states, “The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor.”

In other words, the more we use high-stakes tests to assess students, teachers, schools, and systems the corruptions and distortions that inevitably appear compromise the construct validity of the test and make scores uninterpretable. It is not difficult to imagine the flow-on effects of high stakes testing in developing countries already fighting the scourge of corruption in their education systems. The [Special Issue](#) of the journal *Assessment in Education: Principles, Policy and Practice*, Volume 19, Number 1, 2012 contains a review of the consequences of high stakes testing in developing countries as well as in Australia.

Is it too radical to suggest such tests should be done away with altogether? Perhaps. But, in the real world, calls for data and pressures for testing will persist and educationists and psychometricians will have to address these. Equally, those who like the idea of international comparative testing have to address concerns with the cross-cultural validity of indicators, the consequences of the testing they have in mind, and how these consequences will be managed. With our current state of knowledge, it is dangerous to think that the use of these high stakes tests can have a constructive role in the development of education. Just as Beatty and Pritchett point out that setting overambitious learning goals *may* be counterproductive, the evidence is accumulating that the use of high stakes indicators *will* be counterproductive.

In a subsequent blog, future directions on these issues will be discussed.

This is the first of a two part series on high stakes testing. The second part can be found [here](#).

[Robert Cannon](#) is an Associate of the Development Policy Centre and is presently working as an evaluation specialist with the USAID funded PRIORITAS Project in Indonesian education.