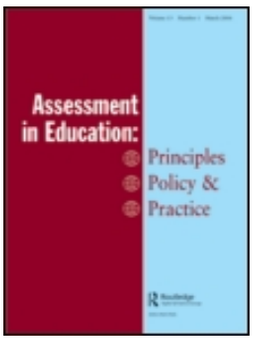

Routledge
Taylor & Francis Group

Assessment in Education: Principles, Policy & Practice

ISSN: 0969-594X (Print) 1465-329X (Online) Journal homepage: http://www.tandfonline.com/loi/caie20

High-stakes testing – value, fairness and consequences

Gordon Stobart & Theo Eggen

To cite this article: Gordon Stobart & Theo Eggen (2012) High-stakes testing – value, fairness and consequences, Assessment in Education: Principles, Policy & Practice, 19:1, 1-6, DOI: 10.1080/0969594X.2012.639191

To link to this article: http://dx.doi.org/10.1080/0969594X.2012.639191

Assessment in Education: Principles, Policy & Practice
Vol. 19, No. 1, February 2012, 1–6

Routledge
Taylor & Francis Group

EDITORIAL

High-stakes testing – value, fairness and consequences

High-stakes testing has been with us for over two thousand years and is steadily increasing both in scale and range. This special issue considers some of the main uses of these tests (a term used loosely to cover many forms of assessment) and their impact. Tests become 'high-stakes' when the results lead to serious consequences for at least one key stakeholder. These consequences could be educational or occupational life chances for individual candidates. This is the case when, for example, testing is used for selection in education and training or to gain credentials that provide 'a licence to practise'. Where tests are used for accountability purposes in evaluating performance and to determine whether targets have been met, they become high-stakes for institutions such as schools and colleges, especially if they affect funding and recruitment. Current international comparisons, for instance the *Programme for International Student Assessment* (PISA), have introduced a new high-stakes phenomenon – tests that are low-stakes for the individuals taking them and for their schools but high-stakes for politicians, policy makers and governments.

While assessments are often used for multiple purposes (Newton 2007), this special issue focuses on three main high-stakes uses: selection and placement, raising standards and accountability. It provides case studies from a range of countries on the impact of high-stakes tests, many of them recently introduced to make selection fairer or to improve educational standards and accountability. Alongside these studies we include two papers which look at technical aspects, one examining models of validation in complex high-stakes assessment contexts and the other examining the reliability of the awarding and certification process in a high-stakes school graduation diploma.

Selection, placement and certification

Historically these have been the original, and dominant, uses of high-stakes assessment. Much of the selection was for occupational progression. Its origins are found in the Chinese Imperial Examination System for civil service selection, which goes back to the competitive examinations during the Han dynasty two thousand years ago (see Carless 2011, for a fuller account). At the heart of this was the attempt to provide fairer selection for schooling, government positions and the military than the nepotism that prevailed (though women and many males in manual work were excluded). By the time of the Ming dynasty (1368–1644 AD) the examination system was fully developed and extremely rigorous, including being locked in an examination cell for three days, the candidates' scripts being copied so that handwriting would not be recognised, and double marking.

ISSN 0969-594X print/ISSN 1465-329X online

http://dx.doi.org/10.1080/0969594X.2012.639191
http://www.tandfonline.com

Like all competitive assessments in which the results affect life chances, there is the risk of seeking to gain advantages over other candidates. Like its modern counterparts, the Chinese system had to reckon with the development of a coaching industry, which would advantage the more affluent, and with cheating. The predictability of the examinations made them a target for smuggling in answers, countered by robust body searches. The cells were exposed so that behaviour could be observed (and the rain come in), but still there were examples of the use of tunnelling. Collusion between test-takers and officials led to severe penalties (in 1657 it led to seven officials being given the death sentence).

We recount these examples because these issues remain with us today and feature in several of the articles. George Bethell and Algirdas Zabulionis provide us with a fascinating account of how high-stakes university selection examinations have been introduced in former Soviet Republics to make selection fairer and to move away from political nepotism. In some of these countries corruption remains part of the social context, so what steps have to be taken to keep the process fair and reliable? What we find are modern equivalents of the Chinese system, with CCTV replacing surveillance towers, examiners being kept away from the public until the exam is taken and scripts being marked anonymously through the use of sophisticated digital technology. The stakes in all this are extremely high: Bethell and Zubulionis observe that 'a single mark can make the difference between, for example, a university place and a year in military service' (p. 17), and so security and reliability become paramount – sometimes with costs to construct validity.

Iasonas Lamprianou's country profile of Cyprus focuses on the unintended consequences of the rapid implementation of a new high-stakes university selection examination, which was introduced at short notice because of a European Union decision. When policy makers sought to combine it with the school graduation examination, so that it had a dual purpose, a number of unintended consequences followed. Because of its high-stakes selection role the new examination has become a very public, and political, concern. Lamprianou also examines how the intention to reduce the influence of the private exam preparation industry, which has traditionally advantaged the more prosperous, has fared.

A third case study of the impact of high-stakes selection examinations is Jerome De Lisle, Peter Smith, Carol Keller and Vena Jules' analysis of the outcomes of the 11+ selection examinations in Trinidad and Tobago. In terms of life chances, the selection examination for secondary schooling could historically claim to carry the highest stakes for individuals. When secondary schooling was rationed it meant the difference between finishing formal education and gaining all the opportunities that came with additional schooling. In an age of universal secondary education its role in many countries is placement as it may determine which educational track students will enter or at which school they get accepted. These are powerful drivers across the world, with parents desperate to get their children into prestigious schools. Again this has led to a coaching industry as parents seek to maximise their children's chances.

The fairness and reliability of secondary selection tests has always been a concern (Gardner and Cowan 2005) given the impact on life chances. This is particularly the case when the outcome rests on a single result from a single assessment. Bourdieu (1991) observed: 'between the last person to pass and the first person to fail, the competitive examination creates differences of all or nothing that can last a lifetime' (120). De Lisle et al.'s analysis looks at some of the equity issues that

could undermine the validity of the 11+ assessment, focusing on gender, geography and assessment design. Assessment design includes what is included and how changing this, to improve equity, has social consequences.

Setting and raising standards

The use of assessments to evaluate and improve the performance of schools, colleges and training institutions is a widely recognised, and very public, purpose. While it may have a very contemporary feel, there are plenty of historical precedents (see Stobart 2008). Twice-yearly written examinations were introduced at Cambridge University at the end of the eighteenth century to improve the performance for its students. The use of external written examinations to raise standards then percolated down to schools, leading in England to the development of the university examination boards which set school examinations and used them for selection as admissions to university became more open.

The use of high-stakes testing for school accountability is exemplified in the United States through the No Child Left Behind (NCLB) legislation, with its financial consequences for schools and teachers. This 'incentive' approach is not without its history. 'Payment-by-results' was introduced in England through the 1862 Revised Code at a time of increasing demand for elementary schooling. The scheme introduced grants for schools, which directly affected teachers' salaries, based on the performance of pupils in reading, writing and arithmetic tests. The assessments were conducted by visiting school inspectors. Like the current NCLB legislation, the intentions were good; teachers would have to prepare all their pupils, not just favour the higher-achieving ones.

Like other high-stakes tests the consequences of payment-by-results, a scheme that continued for 30 years, were mixed. The main negative impact was how it affected teaching and learning, which soon became focused on drilling for the tests. In a scathing indictment of its effects on learning a Chief Inspector commented:

> The children …were drilled in the contents of those books until they knew them almost by heart. In arithmetic they worked abstract sums, in obedience to formal rules, day after day, and month after month; and they were put up to various tricks and dodges which would, it was hoped, enable them to know by what precise rules the various questions on the arithmetic card were to be answered… Not a thought was given, except in a small minority of schools, to the real training of the child, to the fostering of his mental (and other) growth. To get him through the yearly examination by hook or by crook was the one concern of the teacher… To leave a child to find out anything for himself, to work out anything for himself, would have been regarded as proof of incapacity, not to say insanity, on the part of the teacher, and would have led to results which, from the 'percentage' point of view, would probably have been disastrous. (Holmes 1911, 107–8)

Research into current accountability testing suggests that similar risks are still with us. While there may be positive consequences in terms of teachers working harder and more effectively to cover more material (Koretz, McCaffrey, and Hamilton 2001), this may also restrict the curriculum to those subjects that will be tested (Boyle and Bragg 2006) with an emphasis on coaching to the test (Beverton et al. 2005). The most negative consequence would be cheating, either through directly aiding students or through 'playing the system' by manipulating entries, for example by retaining students in the year below the test year or encouraging them to drop out (Hursh 2005).

A recent major US review conducted by the National Academies of Sciences (Hout and Elliot 2011) on the impact of incentives and test-based accountability in education reports similar mixed benefits. Their main conclusion is:

> (1) Test-based incentive programs … have not increased student achievement enough to bring the United States close to the levels of the highest achieving countries. When evaluated using relevant low-stakes tests, which are less likely to be inflated by the incentives themselves, the overall effects on achievement tend to be small and are effectively zero for a number of programs. (S-3)

Despite this kind of evidence, policy makers are still drawn to high-stakes tests which offer relatively simple accountability measures and allow comparisons to be made between schools and local administrations. One such recent development is the introduction of national tests in Australia, a country which previously has only had state-based assessment. Val Klenowski and Claire Wyatt-Smith provide a telling account of the impact of the introduction of the National Assessment Program – Literacy and Numeracy (NAPLAN), which includes school 'league tables' which have not previously been available. Given what we know about the consequences of high-stakes accountability testing, has Australia learned any lessons that would help to mitigate negative impacts?

When low-stakes become high-stakes: the impact of international comparative studies

A more recent high-stakes phenomenon has been the way in which international comparative studies such as *Trends in International Mathematics and Science Study* (TIMSS), *The Progress in International Reading Literacy Study* (PIRLS) and PISA have become high-stakes. Here the consequences are not for the students or schools, since the sampling methodologies mean no direct consequences for them as they are not identified in the results. The consequences are for politicians and policy makers who have to respond to their country's position in the league tables that become the focus of public and policy concern. This is particularly the case when a country does worse than expected, either by sliding down the league table or by doing worse than neighbouring countries. An example of this would be the 'PISA Shock' experienced by Norway in the 2000 PISA study. When a country that has one of the highest per capita investments in education scored below the PISA average and was ranked below its fellow Scandinavian countries, there were extensive political repercussions as the opposition seized on these findings (Baird et al. 2011). This also led to a programme of educational and assessment reform, as it has in other countries, for example Germany and Denmark.

Sarah Howie's study of the impact of South Africa participating, for the first time and as the first African country, in the 1995 TIMSS study and then, as one of two African countries, in the 2006 PIRLS study provides a powerful case study of the impact of taking part in such studies. As a developing nation seeking to overcome its historical legacy, this was a brave commitment to monitor standards. The poor results had a major political and policy impact, as did deciding not to take part in the 2007 TIMSS study. It has led to the setting up of other national monitoring approaches that were seen as more constructive in how they reflect what is very slow progress in the struggle to raise standards.

The quality of high-stakes testing

The consequences of high-stakes testing mean that the quality of the testing instruments, the awarding procedures and the valid interpretation of the results have to be of the highest quality. We see in several of the articles how those responsible for the tests seek to make them as fair as possible for the candidates. The task is to optimise what can be assessed (construct validity), how best it can be assessed (fitness-for-purpose) and how reliability can best be ensured. These raise important theoretical and technical issues.

Two of the articles directly address the theoretical issues around validity and reliability in high-stake assessments. Martha Koch and Christopher DeLuca argue that the multiple purposes that high-stake tests are used for require a re-thinking of conventional validity theorising. Much validity theorising focuses on the use of a single instrument in relation to a specific purpose (for example, Crooks, Kane, and Cohen 1999). When there are multiple purposes, the approach is to validate each one separately. Koch and DeLuca argue that this approach does not do justice to the interactions between these purposes. They propose a model of validation that addresses this, using narrative case description as a better representation of the complexity of large-scale assessment systems. They demonstrate this approach in relation to Ontario's Grade 9 Mathematics assessment. In an era when assessment is regularly used for multiple purposes, this article offers important new thinking on validation.

The reliability of awarding procedures when results of several tests are combined is the focus of the article by Peter van Rijn, Anton Béguin, and Huub Verstralen. Their case study is of another high-stakes assessment, the Dutch secondary school leaving diploma. This is awarded on a pass/fail basis but represents the combination of examination results and teacher assessments in a variety of subjects. How can these diverse results be combined in a way that reduces the risk of misclassifying candidates in the final pass/fail result? The authors consider a variety of technical approaches drawing on test theory and model the alternatives to establish which are the most appropriate decision rules for aggregating results. The article serves as a valuable reminder of the importance of establishing valid assessment procedures, especially when the results generated are of such importance to students' life chances.

These articles are complemented by Gordon Stanley's review of *Secondary School External Examination Systems* (2009), edited by Vlaardingerbroek and Taylor. From the 16 case studies of different countries Stanley identifies some of the main themes common to these end-of-secondary selection examinations. These include a concern with standards as universal secondary education sees an increasing proportion of the cohort taking selective examinations, which in turn leads to concerns about falling standards. Fairness is another major theme, with the pressure in many countries to move to machine-markable formats in order to increase reliability.

We hope that these articles will provide new and relevant evidence about the impact, both positive and negative, of high-stakes testing as well as addressing some of the theoretical and technical issues involved in providing quality high-stakes assessments.

References

Baird, J., T. Isaacs, S. Johnson, G. Stobart, G. Yu, T. Sprague, and R. Daugherty. 2001. *The policy effects of PISA*. Oxford: Oxford University Centre for Educational Assessment.

Beverton, S., T. Harris, F. Gallannaugh, and D. Galloway. 2005. *Teaching approaches to promote consistent level 4 performance in key stage 2 English and Mathematics*. DfES Research Brief 699. London: DfES.

Bourdieu, P. 1991. *Language and symbolic power*. Cambridge, MA: Harvard University Press.

Boyle, B., and J. Bragg. 2006. A curriculum without foundation. *British Educational Research Journal* 32, no. 4: 569–82.

Carless, D. 2011. *From testing to productive student learning: Implementing formative assessment in Confucian-heritage settings*. Abingdon: Routledge.

Gardner, J., and P. Cowan. 2005. The fallibility of high stakes '11-plus' testing in Northern Ireland. *Assessment in Education: Principles, Policy & Practice* 12, no. 2: 145–65.

Holmes, E.G.A. 1911. *What is and what might be: A study of education in general and elementary education in particular*. London: Constable & Co.

Hout, M., and S.W. Elliot. 2011. Incentives and test-based accountability in education. National Academies Press. http://www.nap.edu/catalog.php?record_id=12521 (accessed December 1, 2011).

Hursh, D. 2005. The growth of high-stakes testing in the USA: Accountability, markets and the decline of educational equality. *British Educational Research Journal* 31, no. 5: 605–22.

Koretz, D., D. McCaffrey, and L. Hamilton. 2001. *Towards a framework for validating gains under high-stakes conditions*. CSE Technical Report 551. Los Angeles, CA: CRESST/RAND Education. http://www.cse.ucla.edu/products/reports/TR551.pdf (accessed December 1, 2011).

Newton, P.E. 2007. Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy & Practice* 14, no. 2: 149–70.

Stobart, G. 2008. *Testing times: The uses and abuses of assessment*. Abingdon: Routledge.

Gordon Stobart
Institute of Education, University of London, UK

Theo Eggen
CITO/University of Twente, The Netherlands