

Policy Analysis and Program Evaluation

Policy Analysis and Program Evaluation

Australian Council for Educational Research

Year 2008

Output Measurement in Education

Andrew Dowling
ACER

This paper is posted at ACEReSearch.
http://research.acer.edu.au/policy_analysis_misc/2

Output measurement in EDUCATION

December 2008

by **Andrew Dowling**
Policy Analysis and Program Evaluation Unit

Contents

Introduction	1
What is output measurement and why is it here?.....	2
Key features of output measurement systems	2
What other countries are doing.....	5
Opposing views on output measures and performance	6
Lessons for Australia.....	7
Conclusion.....	8
Works cited	9

Tables and figures

Figure 1: Spending and outcomes in the OECD.....	2
Figure 2: Timeline of national student testing in Australia.....	4
Table 1: OECD country means: Use of comparative assessments.....	5
Box 1: Student testing regimes in high-performing OECD countries.....	5

Introduction

Governments can no longer justify their performance in education in terms of inputs; that is, in terms of the amount of new money they have provided, or the number of new teachers they have employed, or the range of new computers they have installed. It has been observed that 'today, educators need to show how they have transformed current and new dollars into student achievement results, or the argument that education needs more - or even the current level of - money will be unlikely to attract public or political support' (Odden and Picus, 2008, p. 26). Output measures, particularly those related to student achievement, are the new bottom line in education.

The emphasis on accountability through external testing is driven by the growing realisation that education is a major factor in economic development and the consequent understanding that it is the quality of education that is most important (Hanushek and Wößmann, 2007). Accountability for quality has been given a harder edge, often in the face of opposition from the education profession, through standardised tests of cognitive skills (Popham, 2003).

The essay provides an overview of

- The development of output measurement;
- The extent to which such measures have been used in education systems to improve accountability;
- Evidence of their effectiveness; and,
- Implications for Australia.

The essay argues that performance measures constitute a positive shift in education but they haven't gone far enough. More work needs to be done in evaluating the programs that are meant to

improve student performance. The programs that are designed for the most disadvantaged students often escape any systematic form of evaluation yet systems need to formally identify what actually works, and doesn't work, in schools.

What is output measurement and why is it here?

Accountability systems have been defined as those that 'combine clear standards, external monitoring of results, and corresponding rewards and sanctions based on performance indicators' (OECD, 2007a, p. 9).

The rise of accountability in education is due primarily to the very significant investments made into education. A recent McKinsey report found that despite 'massive' spending on education by the world's governments, totalling \$2 trillion in 2006, performance has barely improved in decades (McKinsey and Company, 2007). Other research has come to similar findings (Hanushek, 1997; Hanushek & Rivkin, 1997; Hanushek and Wößmann, 2007; Pritchett, 2003; Odden and Picus, 2008; Leigh and Ryan, 2008). Pritchett reproduces findings by Gundlach, Wößmann & Gmelin (2000) that over the period 1970-94, nearly every OECD country witnessed an enormous expansion in expenditures per pupil, while their maths and science performance either flat-lined or deteriorated (see Figure 1).

The fact that funding does not often correlate with performance is a reason for the focus on outputs in education. Outputs can be defined as an individual's, school's, or nation's performance, as

measured by standardised tests. A standardised test is one where the method of administering the test, including the test conditions and system of scoring, is regulated and controlled so that it is consistently applied across multiple groups. The purpose of standardised tests is to better judge achievement by relating performance (whether it be by the student, teacher, school, or nation), to a wider population.

Output measures have been used in the past to criticise education systems and will continue to be used for this purpose. Further, the relationship between funding and output measures has been the subject of heated academic debate (see, for example, Hanushek, 1996, and Greenwald, Hedges and Laine, 1996a & 1996b). But output measures are also an extremely powerful rationale to continue justifying increased spending on education.

Key features of output measurement systems

The two main features that distinguish output measurement systems are whether:

- a) They have penalties attached or not; and whether;
- b) They are national in scope or not.

The United States (US) is an example of a system that has penalties attached but is not nationally organised, while Australia's system is national but does not lead to any specific penalties.

Assessments with penalties attached are often referred to as 'high stakes.' This term should probably be confined to instances where tests are truly 'high stakes,' such as in exit school examinations

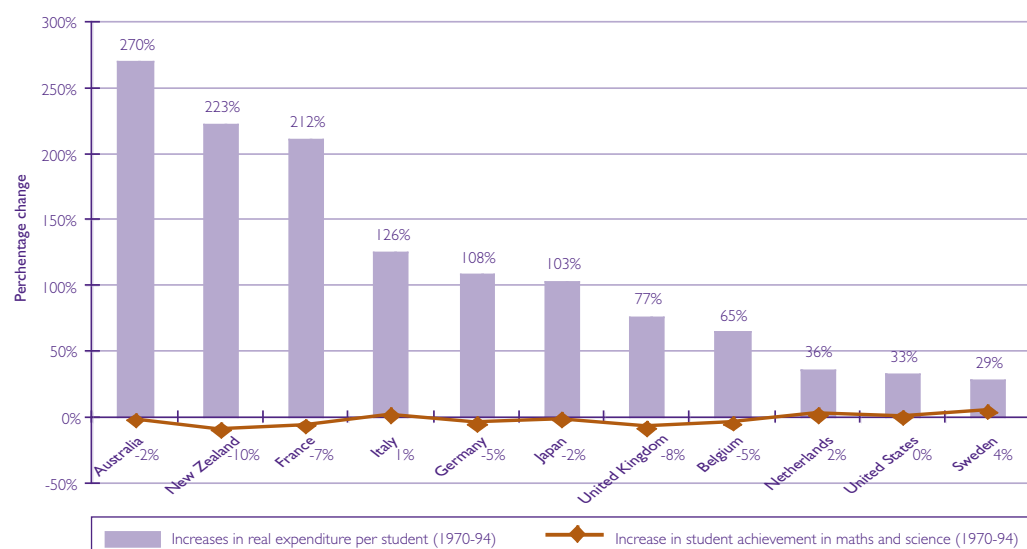


Figure 1: Spending and outcomes in the OECD

Source: Pritchett (2003), adapted from Gundlach, Wößmann & Gmelin (2000). This data are also reproduced in the McKinsey report (2007).

Note: Student achievement data comes from TIMSS (*Trends in International Mathematics and Science Study*).

or medical entrance exams, both of which have immediate consequences for the individuals who sit them. The type of assessments we are referring to are a form of 'standards test,' quality control systems designed to keep schools, and school systems on their toes rather than being 'high stakes' for the individuals who complete them.

The United States

The US example illustrates how the emphasis on education measurement combines with a faith in the free market. The efficient operation of any market requires good information and this is exactly what student testing provides. The idea that market forces can advance society much more effectively than government intervention is, in fact, one of the major reasons behind the introduction of student testing on a large scale.

In the US, standardised achievement tests have been designed to facilitate a market in education services by increasing competition and choice. The US *No Child Left Behind (NCLB) Act* was introduced to Congress in 2001 and signed into law by President Bush in January 2002 (NCLB, 2002). Colloquially referred to as the *No Child Left Untested Act*, this law encourages students to move schools and for schools to be restructured as a consequence of continued poor performance in testing. Schools not making progress face 'increasingly rigorous sanctions designed to bring about meaningful change,' ranging from supporting students to transfer to other public schools to restructuring schools (US Department of Education, 2002, p. 17). Thus the description, 'high stakes,' which, as mentioned above, is probably a misnomer:

The language that inaugurated the NCLB Act is almost exactly the same as that which heralded the start of national student testing in Australia, both of which occurred in 2001. Dr David Kemp, Australia's Education Minister at the time, observed that, 'this agenda is all about parents' rights to have objective standards against which they can compare their child's and their school's performance' (Kemp, 2000) while the NCLB Act was designed, 'so that students, teachers, parents, and administrators can measure progress against common expectations for student academic achievement' (NCLB Act, Sec 1001, paragraph 1). Information was the crucial issue in both cases, linked in both cases to a desire for a more open market in education.

Conventional wisdom in the US is that the NCLB Act is 'on target' but experts in educational measurement note that while it has inaugurated a 'testing revolution,' the law is based on 'the nearly unchallenged belief, with very little supporting evidence, that high-stakes testing can and will lead to improved education': 'Apparently, most policy-makers assume that accountability in education can be accomplished only through the imposition of

high-stakes testing, although there is no compelling body of evidence to support that assumption' (Brennan, 2006, p. 10).

It remains to be seen whether the US experience is an aberration or a harbinger of change. The US Department of Education points to significant, quantifiable gains resulting from the NCLB Act (US Department of Education, 2007), while the former president of the American Educational Research Association, David Berliner, believes these gains to be illusory:

If the intended goal of high-stakes testing policy is to increase student learning, then that policy is not working. While a state's high-stakes test may show increased scores, there is little support in these data that such increases are anything but the result of test preparation and/or the exclusion of students from the testing process (Amrien and Berliner, 2002).

Australia

Australia's system of measuring student performance has a unified, national scope although Australia's non-financial school data (the subject of this essay) is much more organised than its financial data (see Dowling, 2008).

Figure 2 shows a time-line for the introduction of national testing in Australia. What becomes immediately apparent is that politics and technology are closely linked in this chronology. The viability of national testing in Australia was dependent on the States and Territories maintaining ownership of the testing process, which was facilitated by the Rasch model of measurement that helped different tests be equated so that national data could be derived.¹ Moreover, Item Response Theory (IRT), of which the Rasch model is a part, had only been readily accessible, from a practical point of view, for about a quarter of a century, with the introduction of relatively fast microcomputers in the 1980s (Brennan, 2006). But if Australia shows that student testing is a product of its time, it remains to be seen whether standards testing has come of age.

The question arises as to whether Australia will attach penalties to its national testing, and whether this development is inevitable. Australia currently does not have the same penalties attached to student testing as the US but the architecture is in place, to a greater degree than in the US, for individual schools to be compared on national tests. Part of the answer to Australia's direction lies in what other high performing countries are doing in this area.

¹ Georg Rasch (1901 - 1980), a Danish mathematician, statistician, and psychometrician, created models that allowed items from different tests to be equated onto a common measurement scale. This in turn meant that test equating was more feasible and defensible (Sadeghi, 2006, p. 2 & 8).

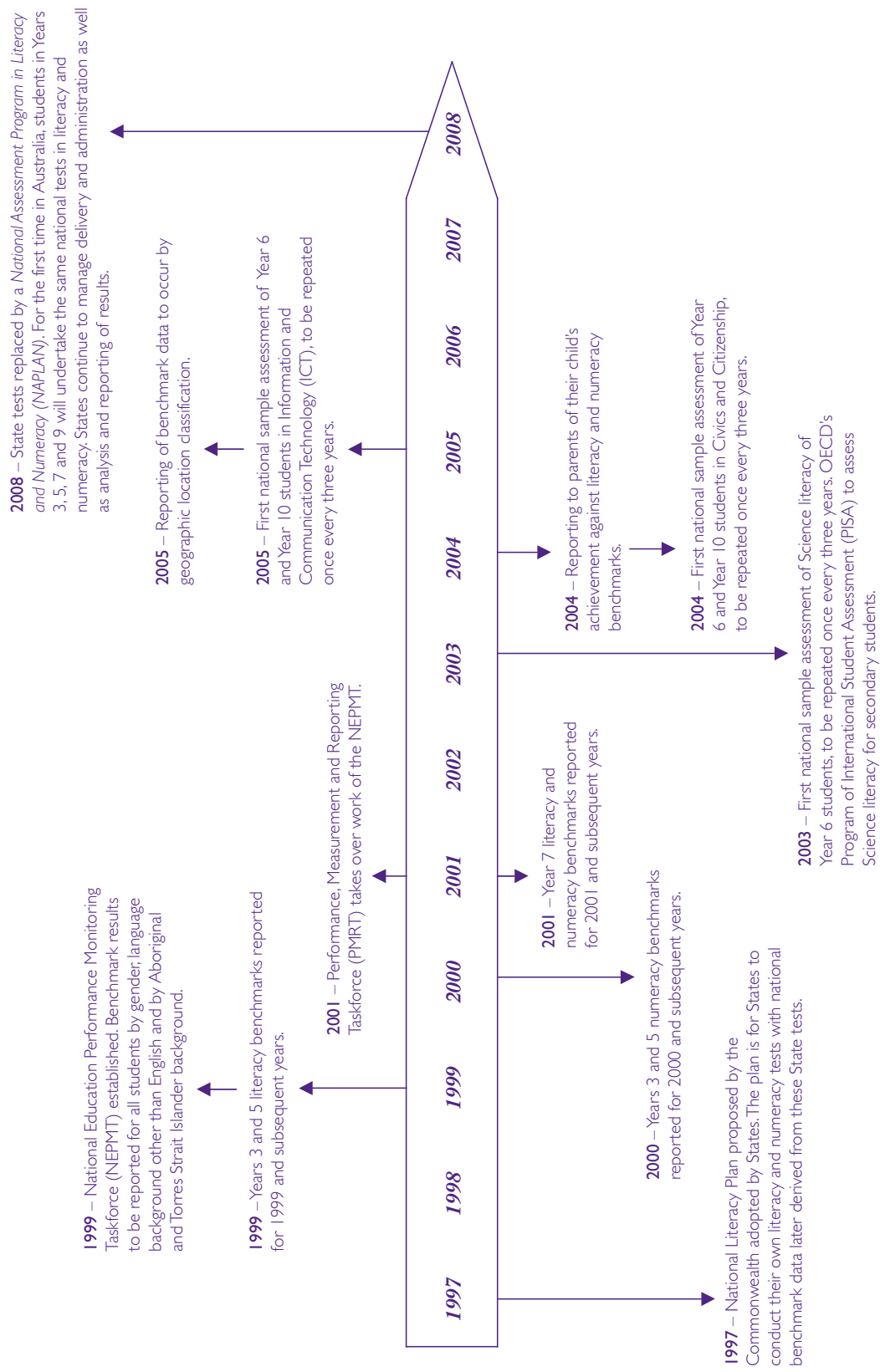


Figure 2: Timeline of national student testing in Australia

What other countries are doing

Output measures that compare schools with each other and with national averages are surprisingly under-developed in high performing OECD countries. Two related OECD studies recently correlated all features of accountability, autonomy and choice at the country level with the 2003 Program of International Student Assessment (PISA). A 'performance study' correlated achievement data while an 'equity study' correlated relevant data from hundreds of thousands of students from various OECD countries (OECD, 2007a & OECD, 2007b). In compiling this study, the school background questionnaires from the 2003 PISA were used to construct country aggregate means of accountability, which are reproduced in Table 1.

Table 1: OECD country means: use of comparative assessments (in descending order)

Country	Assessments for:	
	Comparing to district and nation	Comparing to other schools
United States	0.91	0.80
United Kingdom	0.89	0.84
New Zealand	0.87	0.74
Hungary	0.86	0.77
Iceland	0.84	0.66
Sweden	0.73	0.65
Poland	0.71	0.62
Canada	0.70	0.53
Norway	0.64	0.47
Netherlands	0.63	0.47
Korea	0.62	0.55
Turkey	0.59	n/a
Finland	0.56	0.35
Australia	0.55	0.39
Mexico	0.55	n/a
Czech Republic	0.50	0.55
Slovak Republic	0.46	0.48
Italy	0.33	0.29
Portugal	0.33	0.22
Luxembourg	0.22	0.10
Germany	0.21	0.17
Switzerland	0.19	0.16
Spain	0.18	0.17
Japan	0.18	0.12
Ireland	0.17	0.09
Austria	0.12	0.38
Greece	0.12	0.16
Belgium	0.10	0.07
Denmark	0.06	0.03

Source: OECD, 2007a & b, Table A.2 (Appendix A.3).

The consistently best performing OECD countries on PISA (Finland, Japan, the Netherlands and Korea; Chinese Taipei not represented in the list) are clustered in the middle of the group, with Japan near the bottom. These countries' high academic performance is clearly not matched by their willingness to compare schools to district or national performance, or with each other.

What is also surprising is that none of the top performing OECD countries (Finland, Japan, the Netherlands, Korea, Chinese Taipei) have any form of national assessment, certainly none that compares to Australia. The situation in each of these countries is summarised in Box 1 below, by alphabetical order of country:

Box 1: Student testing regimes in high-performing OECD countries

Chinese Taipei

In Chinese Taipei, there are no national assessments of student progress for accountability purposes although recently, national admission tests have been introduced. Since 2006, all Taiwanese junior-high students (aged 14 - 15 years) have to take the Basic Competence Test (BCT) held twice each year. A BCT has also been introduced for sixth grade students (aged 11 - 12 years) in Chinese and Mathematics (this test also asks students to identify the amount of TV watched every day and the amount of daily computer usage time). The purpose of the year 6 test results is to act as a reference only to teaching practices and is not made available to the public (Chang, Lee, and Yeh, 2006).

Finland

Finland leads the world in literacy and numeracy yet it has no large-scale testing programs in its elementary schools. In the 1990s, Finland abandoned uniformity in curriculum content and moved to basing their teaching and learning on curriculum standards while allowing schools flexibility in the content of the curriculum in achieving these standards (Ministry of Education, 2006).

Japan

In 2007, the Japanese Education Ministry, through the National Institute for Educational Research (NIER), conducted its first national survey of school academic achievement in 43 years. A Nationwide Academic Ability Assessment (NAAA) is now administered to students in the final year of primary education (11 - 12 years of age) and in the final year of lower secondary school (14 - 15 years of age). These tests assess reading, writing and maths, and also ask about

students' eagerness to learn and their daily life habits, including questions such as how many hours they study at home and whether they eat breakfast every morning. Test results are not publicly announced. Instead, local governments and schools receive information on the results. Schools can then determine their position by comparing the national averages, which will be announced by the Government. Students are informed of their results (Andrews, C. et al, 2007).

Korea

Korea does not have a national assessment of student progress for accountability purposes but does have a national sample of student achievement, the principal aim of which is to monitor the curriculum. Small samples of students (0.5 to one per cent of the whole student population) in Years 6 (aged 11-12 years), Year 9 (aged 14-15 years) and Year 10 (aged 15-16 years) are involved in the assessments and two subjects are assessed each year, usually on a rotating basis. Korea has recently moved to a formal written test rather than multiple choice assessments for these national sample tests (KICE, 2007 and Andrews, C. et al, 2007).

The Netherlands

The Netherlands do not appear to have national assessment of student progress for accountability purposes. There is national assessment conducted once every five years in the final year of primary school, when students are 12 years of age (known as CITO tests), which relate students' achievement to the main objectives of primary education (CITO, 2006). There is also a compulsory test at 15 years of age but this is only intended to help guide students' progression to the appropriate school and course type (Andrews, C. et al, 2007).

The United Kingdom

The UK has developed national student tests for accountability purposes from a very early stage. A 'Foundation Stage Profile' (to be replaced with an 'Early Years Foundation Stage Profile' in 2008) assesses children's progress and learning needs from age three to the end of the academic year in which a child has their fifth birthday (Andrews, C. et al, 2007).

In regard to national school assessments, all students in maintained (publicly funded) schools (and some in private, independent schools), at the ages of 7, 11 and 14 are assessed via National Curriculum Assessment, the purpose of which is to improve teaching and learning and provide

information for parents and the public to help them judge the quality of the education being provided. Independent (private) schools are encouraged, but not required, to take part in these statutory assessments. Statutory assessments involve externally set and marked tests which have so far focused on English, mathematics and science.

If there is a link between output testing and performance, one would have thought that advanced directions in education measurement would be most likely evident in countries at the forefront of educational performance and improvement (assuming, of course, that academic results say something about education systems). But as the information in Box 1 makes clear, this is not the case.

In Box 1, the country with the most developed forms of national assessment, particularly for accountability purposes, is the UK, even though it is the lowest performing country amongst this group of very high performing countries as measured on PISA tests. There are three possible explanations for this phenomenon:

- a) National testing for accountability purposes does not improve student performance;
- b) National testing for accountability purposes does improve student performance and, if introduced, would lift the performance of high performing countries even higher; or
- c) Testing for accountability decreases as performance increases, as there is a decreased need to monitor performance.

Of course, there are many reasons behind the superiority of Finland, Japan, Korea, Chinese Taipei and the Netherlands on PISA tests that have nothing to do with educational measurement. These reasons include cohesive social structures and relative cultural homogeneity. But it would be interesting if these countries' performance improved if accountability requirements, based on national tests of cognitive skills, were also introduced on a wider basis.

Opposing views on output measures and performance

The OECD studies mentioned above found that although the highest performing OECD countries only moderately use comparative tests, all types of accountability systems were, in general, effective, whether they were aimed at the student, teacher, or the school. Although the OECD authors advised caution in interpreting their school accountability results, the result was that students perform better when their schools use assessments to compare themselves to district or national performance (OECD, 2007a, p. 29).

These findings contradict previous research, which has found that a jurisdictional emphasis on testing for accountability purposes was generally ineffective. For example, an influential study of the performance of fifty states in America found that states that developed extensive testing systems coupled with rewards and sanctions failed to improve student performance, according to US National Assessment of Educational Progress (NAEP) longitudinal data, while states that invested heavily in teacher education and standards did improve (Darling-Hammond, 2000). The recent OECD study contradicts this finding. In fact, the OECD performance study found that testing for accountability, combined with autonomy and choice for schools, produce students who 'perform substantially better on cognitive skills in mathematics, science and reading as tested in PISA 2003 than do students in school systems with less accountability, autonomy, and choice' (OECD, 2007a, p. 58).

The OECD researchers explained this effect as due to better alignment between principals and agents. One example of a principal-agent relationship in education is when a principal (e.g., the parent) commissions an agent (e.g., the head of a school) to perform a service (the education of the child) on her behalf. Another example is when a government (the principal) commissions an education authority (the agent) to improve school results (the service) for a given state. In both cases, incentives can be introduced to make the agent do what the principal wants, particularly if the agent's interests differ from that of the principal.

Counter-arguments

The problem noted by many educators is that the agent may well do what the principal wants, but at the expense of a good education. This is essentially what David Berliner says the NCLB Act is doing: increasing scores by narrowing focus. Others have gone further, stating that the US accountability regimes create perverse incentives, such as 'curricular reductionism, excessive test-focused drilling, and the modelling of dishonesty [where teachers act fraudulently to increase test scores]' (Popham, 2003, p.12). The claim of dishonesty is levelled system-wide, with the claim that actors in accountability systems collude in meeting specified targets so that the targets eventually 'bear as much likeness to reality as did the production goals of the former USSR' (Mortimore, 2008). There is a widespread belief that these accountability systems, at the very least, force teachers to teach to the test: 'The notion that testing limits the nature of teaching is pervasive' (Pellegrino, 2004, p.8).

The putative loss of a better, wider education remains at the level of anecdote precisely because it cannot be measured. But its relevance can be

gauged by explanations offered for why money appears to have such a low impact on student performance. One reason given is that most of the extra money has been spent on non-core subjects (such as art, music, physical education, drama, health, vocational education, etc) and students with special needs, precisely those subjects and students who are not assessed through standardised tests (Odden & Picus, 2008, p. 184). The notion that scores are related to what is spent suggests that educational measurement may construct new values in the classroom.

The common view that testing is not the same as learning (and may in fact be harmed by excessive testing), has no empirical basis; yet is supported by economic explanations for the low impact money has on student performance (namely, that the money is not focused narrowly enough). However, if studies such as those produced by the OECD continue to find performance improvement through comparative output measures, then the use of these systems will increase. In this context, more definitive research on the US experience will be crucial.

Lessons for Australia

It is unlikely that the Australian system will attach penalties to its assessment regime in the near future. The fact that many high performing countries do not do this and a large proportion of the education community is opposed to it would seem to settle the matter. But if more and more countries do take this path and if technological developments allow school and teacher effects to be more precisely identified, then the pressure will grow for Australia to move in this direction. In this context, the development of value-added assessment may be important.

Value-added assessment is a trend that has come from within the education sector, largely in response to that sector's resistance to other forms of accountability systems. If any accountability system is to be imposed on education, most educators would prefer it to be one that isolates their effects. This is what value-added assessment promises to do.

Value-added assessment focuses on a student's growth over a given period of time rather than the absolute levels they attain at a point in time. Theoretically, growth reveals the effects of schools and teachers while achievement does not. However, there are significant problems with value-added assessments, including:

- Most value-added approaches remain highly technical.
- Creating vertical scales is not only statistically challenging, but may introduce more error in longitudinal analysis.

- Missing data on student performance, as well as data linking students to teachers, may become a significant problem as large proportions of students transfer among schools every year.
- It is unclear whether the estimate obtained from a value-added model could be called a teacher or school effect, when all the other factors that influence a student's score are taken into account.

(Rand Corporation, 2004 & Doran and Fleischman, 2005).

In 2004, the Rand Corporation advised that 'the current research base is insufficient to support the use of value-added modelling for high stakes decisions.' But value-added software programs are becoming more widely available, even if implementing these models remains complex (Doran and Fleischman, 2005). The fact that Australia's new national assessment program will continue to use the Rasch model for both vertical and horizontal equating suggests the eventual arrival of value-added assessments, despite the implementation problems.² As this occurs, technological developments that isolate the effect of individual schools and teachers on student performance will only increase the pressure to use these measures for accountability purposes.

Testing may also revert to its tradition role as a diagnostic rather than accountability tool. It has been predicted that the type of mass testing introduced by the NCLB Act and national benchmarking in Australia will eventually be considered a quaint anachronism: 'In 21st century learning environments, decontextualised, drop-in-from-the-sky assessments consisting of isolated tasks and performances will have zero validity as indices of educational attainments' (Pellegrino, 2004). Rather, assessment will become much more targeted at mapping students' knowledge and diagnosing students' misconceptions about specific topics.

The trend towards individual diagnosis matches the laudable move towards 'personalisation' in education, where schools are moving away from Fordist principles of standardised mass production to systems that are fashioned for the individual (Leadbeater, 2004). However, this trend towards the individual would not supplant the equally strong need for increased accountability of systems. In fact, this essay argues that accountability should be even more deeply embedded into education practice. It remains the case that there is generally no culture of measuring program effectiveness at the school

level in Australia, or in most other countries. The practice of benchmarking and public identification of 'better' or 'worse' activities in schools is rarely conducted in any formal way. For example, one of the few evaluations of equity programs in New South Wales public schools proposed a system of continuous monitoring, review and accountability on the assumption that 'it is essential to identify programs that are successful in promoting better outcomes for disadvantaged students' (Lamb & Teese, 2005). However, as one of the report's authors, Stephen Lamb, subsequently noted, it was a continuing problem world-wide that systems simply allocated resources to schools without a clear idea on how they would or should be spent (The Australian, 7 July, 2008). It remains the case that programs and initiatives designed for disadvantaged students frequently escape any systematic scrutiny of their effects.

The reluctance to evaluate also extends to teaching practice. A recent study of the teaching profession found that it 'does not have well-established institutions or procedures for using research to identify and define standards for what its members should know and be able to do - normative structures relating to good practice are weak' (ACER, 2008). Yet educators need to know, in more detail than they do, what works and doesn't work in schools. It would be a positive result if output measures extended further into education, so that school programs were regularly and formally evaluated in terms of their effectiveness.

Conclusion

Educational assessment is not a new concept. China used student tests 3,000 years ago and introduced a national civil service examination system 1,500 years ago, while modern educational test development can be traced to the Industrial Revolution (Oakland et al, 2001, p.4). Yet today's emphasis on output measurement is a new phenomenon, one that can be traced to an evidence-based management philosophy that first transformed Japanese industry after the Second World War and was introduced more broadly to the West in the 1980s.

There is no doubt that the changes inaugurated by output measurement will be profound. This is in contrast to a common view that policy changes in education are invariably superficial and do not affect the reality of school practice. In one striking analogy, such policy changes are likened to a storm on the ocean: 'The surface is agitated and turbulent, while the ocean floor is calm and serene (if a bit murky). Policy churns dramatically, creating the appearance of major changes ... while deep below the surface, life goes on largely uninterrupted' (quoted in McKinsey and Company, 2007). Output measures will disturb

² Horizontal equating places on a common scale tests of the same difficulty while vertical equating places on a common scale tests of different difficulty, usually tests across different year levels, thus allowing longitudinal analysis of individual student performance. Australia's assessment will be both vertical and horizontal in the sense that tests at each grade level are equated from one year to the next.

school life below the surface, mainly because of the deep need for accountability it responds to and the scale of change that is involved.

Output measures are the new currency of an educational market; the new 'bottom line' upon which schools, school systems, and increasingly teachers, will be judged. This essay argues that standardised performance measures should be extended so that equity programs are also evaluated. But the question of whether accountability systems should have penalties attached to them is another matter. Much will depend on authoritative studies of existing initiatives and technological innovations will also be important, particularly value-added models that can isolate the impact of schools and teachers on student performance. However, in either case, the continuing role of standardised assessments in providing reliable information for a new education market is inevitable and justified.

Works Cited

- Amrien, A. L. and Berliner, D.C. (2002). 'High-Stakes Testing, Uncertainty, and Student Learning.' *Education Policy Analysis Archives*, 10:18.
- Andrews, C., Brown, R., and Sargent, C. with O'Donnell, S. (2007). 'Compulsory assessment systems in the INCA countries: Thematic Probe.' Retrieved 30 November 2007, from <http://inca.org.uk>
- Australian Council for Educational Research (ACER). (2008). *Teaching Talent: The Best Teachers for Australia's Classrooms*. Business Council of Australia (BCA) & ACER. Accessed 03 July 2008. http://www.acer.edu.au/documents/MR_080526_BCA08.pdf
- Brennan, R. L. (2006). 'Perspectives on the Evolution and Future of Educational Measurement.' In *Educational Measurement, Fourth Edition*. American Council on Education and Praeger Publishers. pp. 1-16.
- Chang, C-Y, Lee, W-C, and Yeh, T-K. (2006). 'Taiwanese Earth Science Curriculum Guidelines and Their Relationships to the Earth Systems Education of the United States.' *Journal of Geoscience Education*. V54, No.5. November. Accessed 3 December 2007. <http://www.nagt.org/files/nagt/jge/abstracts/chang-v54p620.pdf>
- CITO, The Netherlands Institute for Educational Measurement. (2006) The Dutch Education System. Retrieved 5 December 2007, from, http://www.cito.com/com_about/dutch_edu_syst/eind_fr.htm
- Darling-Hammond, L. (2000). 'Teacher Quality and Student Achievement: a review of state policy evidence.' Centre for the Study of Teaching and Policy, University of Washington. Reprinted in *Education Policy Analysis Archives*, 8:1.
- Doran, H. C. and Fleischman, S. (2005). 'Research Matters: Challenges of Value-Added Assessment.' *Educational Leadership*. Vol 63. No. 3. (November 2005). pp. 85-87.
- Dowling, A. (2008). 'Unhelpfully Complex and Exceedingly Opaque': Australia's School Funding System.' *Australian Journal of Education*. 52.2, August 2008, pp. 129-50.
- Greenwald, R., Hedges, L.V, and Laine, R. D. (1996a). 'The Effect of School Resources on Student Achievement.' *Review of Educational Research*. 66 (3): 361-96.
- Greenwald, R., Hedges, L.V, and Laine, R. D. (1996b). 'Interpreting Research on School Resources and Student Achievement: A Rejoinder to Hanushek.' *Review of Educational Research*. 66 (3): 411-416.
- Gundlach, E. Wößmann, L. and Gmelin, J. (2000). *The Decline of Schooling Productivity in OECD Countries*. Kiel Working Paper No. 926. Kiel Institute of World Economics. April 2000. <http://www.ifw-members.ifw-kiel.de/publications/the-decline-of-schooling-productivity-in-oecd-countries/kap926.pdf> Accessed 19 September 2008.
- Hanushek, E. & Rivkin, S. (1997). 'Understanding the Twentieth-Century Growth in U.S. School Spending'. *Journal of Human Resources*. 32 (1): 35-68.
- Hanushek, E. (1996). 'A More Complete Picture of School Resources Policies.' *Review of Educational Research*. 66 (3): 397-409.
- Hanushek, E. A. (1997). 'Assessing the Effects of School Resources on Student Performance: An Update.' *Educational Evaluation and Policy Analysis*. 19(2): 141-164.
- Hanushek, E. A. and Wößmann, L. (2007). *Education Quality and Economic Growth*. The World Bank, Washington DC. <http://www.gio.gov.tw/taiwan-website/5-gp/yearbook/18Education.htm>
- Kemp, Dr D. (2000). 'Numeracy Benchmarks to be Introduced into Schools.' Media Release K054 Monday, 10 April 2000. Accessed 12 June 2008. http://www.dest.gov.au/archive/ministers/kemp/april00/k054_100400.htm
- KICE, Korea Institute of Curriculum and Evaluation. (2007). Information on Curriculum and Education and Educational Evaluation: National Assessment of Educational Achievement. Retrieved 30 November 2007, from, <http://kice.re.kr/kice/eng/info>
- Lamb, S & Teese, R. (2005). 'Equity programs for government schools in New South Wales: a review.' Centre for Post-compulsory Education and Lifelong Learning. University of Melbourne. December 2005. Accessed 8 July 2008. <https://www.det.nsw.edu.au/media/downloads/research/completedprojects/nswequityrev.pdf>

- Leadbeater, C. (2004). Learning about personalisation: how can we put the learner at the heart of the education system? DEMOS, in partnership with the National College for School Leadership. Accessed 2nd December 2007. <http://www.demos.co.uk/publications/learningaboutpersonalisation>
- Leigh, A, and Ryan, C. (2008). *How Has School Productivity Changed in Australia?* <http://econrssh.anu.edu.au/~aleigh/pdf/SchoolProductivity.pdf> Accessed 17 June 2008.
- McKinsey & Company. (2007). *How the World's Best-Performing School Systems Come Out on Top.* Accessed 17 June 2008. http://www.mckinsey.com/client/service/socialsector/resources/pdf/Worlds_School_Systems_Final.pdf
- Ministry of Education (2006) Education and Science in Finland. Accessed 3 December 2007. <http://www.minedu.fi/>
- Mortimore, P. (2008). 'These protesters are not dinosaurs,' *The Guardian*. Tuesday May 6, 2008.
- No Child Left Behind (NCLB) Act of 2001. Public Law 107-110. 107th Congress. January 8th, 2002. Accessed 12th June 2008. <http://www.ed.gov/policy/elsecd/leg/esea02/pg1.html#sec1001>
- Oakland, T., Poortinga, Y. H., Schlegal, J., and Hambleton, R. K. (2001). 'International Test Commission: Its History, Current Status, and Future Directions.' *International Journal of Testing*, vol 1, pp. 3-32.
- Odden, A. R. and Picus, L. O. (2008). *School Finance: A Policy Perspective, 4th edition.* New York: McGraw Hill.
- OECD. (2007a). *School Accountability, Autonomy, Choice and the Level of Student Achievement: International Evidence from PISA 2003.* December 2007. <http://www.oecd.org/dataoecd/53/59/39839361.pdf> Accessed 10 June 2008.
- OECD. (2007b). *School Accountability, Autonomy, Choice and the Equity of Student Achievement: International Evidence from PISA 2003.* December 2007. <http://www.oecd.org/dataoecd/53/57/39839422.pdf> Accessed 6 June 2008.
- Pellegrino, J. W. (2004). 'The Evolution of Educational Assessment: Considering the Past and Imagining the Future.' Policy Evaluation and Research Center: ETS. Princeton, New Jersey. Accessed 30 November 2007. <http://www.ets.org/Media/Research/pdf/PICANG6.pdf>
- Popham, W. J. (2003). 'Preparing for the Coming Avalanche of Accountability Tests.' *In Spotlight on High-Stakes Testing.* Cambridge, MA. Harvard Education Press. pp.9-15.
- Pritchett, L. (2003). Educational Quality and Costs: A Big Puzzle and Five Possible Pieces. http://ksghome.harvard.edu/~lpritch/edpuzzle_onepiece.doc Accessed 18 June 2008.
- Rand Corporation. (2004). *The Promise and Perils of Using Value-Added Modelling to Measure Teacher Effectiveness.* Rand Education Research Brief. http://www.rand.org/pubs/research_briefs/RB9050/RAND_RB9050.pdf Accessed 12 June 2008.
- Sadeghi, R. (2006). *An investigation of the consequences for students of using different procedures to equate tests as fit to the Rasch model degenerates.* Unpublished Thesis (Ph. D.), University of New South Wales.
- The Australian. (2008). 'No strings school funding 'a failure''; 7 July, 2008. Accessed 8 July 2008 <http://www.theaustralian.news.com.au/story/0,,23979279-2702,00.html>
- U.S. Department of Education. (2002). *No Child Left Behind: A Desktop Reference.* <http://www.ed.gov/admins/lead/account/nclbreferenc/reference.pdf>. Accessed 23rd November 2007
- U.S. Department of Education. (2007). *No Child Left Behind's 5th Anniversary: Keeping Promises And Achieving Results.* <http://www.ed.gov/nclb/overview/importance/nclb5anniversary.html> Accessed 6th June 2008.