Putting the nation to the test, is there room for improvement?

Robyn Timmins
Manager, Assessment Services
Curriculum Corporation

Abstract
This paper examines a sample of items drawn from Australia’s State and Territory tests in the context of an era when assessment data are increasingly used for multiple purposes and scrutinized by much broader audiences than ever before. Test data are now used for national and international comparisons of standards of learning and educational attainment. They are being used by governments and sectors to inform policy priorities for curriculum reform, to improve student learning, pedagogy, school improvement and even in some schools to support the case for a change of leadership.

A stronger focus on public accountability has contributed towards increased efforts to improve the integrity of the data generated and reported to systems and to the Australian government. School leaders and systems are working towards improving the capacity of teachers to interpret data and to use it more effectively to improve student learning and raise standards. In such times of public scrutiny and accountability, what about the test items themselves, how might we raise standards?
In recent times large-scale testing programs in the compulsory years in all Australian States and Territories have assumed a similar level of importance to standards based testing programs in the UK and the US. This has occurred without a national accountability regime similar to America’s No Child Left Behind (NCLB) legislation, which has tied test results to eligibility for categories of federal funding (Resnick, Rothman, Slattery and Vranek 2003-2004). However, in Australia, similar pressures at a local level to meet accountability requirements have pushed testing programs beyond previous boundaries, and have raised the stakes for all parties involved. The pressure to deliver improved results and to put student achievement data to an increasing array of uses brings with it the need for stringent quality assurance processes and public confidence in the quality and integrity of the data. This paper raises some concerns about the quality of test items and suggests that there is room for improvement through shared knowledge and documentation of best practice.

Since the early 1990s and pioneer assessment programs such as Western Australia’s Random Sample Assessment, and early sample testing in each State and Territory to establish Year 3 and 5 benchmark data, our testing programs have gradually been expanding. Local programs have progressed well beyond the initial literacy and numeracy agreements of MCEETYA;

That every child leaving primary school should be numerate, and able to read write and spell at an appropriate level.

That every child commencing school from 1998 will achieve a minimum acceptable literacy and numeracy standard within four years.

(MCEETYA)

Today Australian States and Territories embrace testing in the compulsory years in a wide variety of forms. In the not too distant future we might look back on this expansive testing era as the Golden Age of Assessment.
Testing programs in most States and Territories have grown significantly in the past decade. Around the country, systems have steadily added new year levels, learning areas, practical and non-practical test components, computer based and computer adaptive testing and piloted delivery of on-line services to enhance their suite of testing programs. In 2004, another new initiative is being trialed in Victoria, On-demand Assessment for all year levels P-10.

*This new facility will enable teachers to access calibrated tests from a central assessment test item bank across the curriculum and year levels*

(VCAA 2004).

In addition to the expansion of testing we have added services such as ancillary support programs and resources to assist school leaders and teachers to make better use of student achievement data. For example, the Data Club in Western Australia, on-line data tools in NSW and Victoria, and today almost all states provide a dedicated website with support materials, a helpline and a battery of administrative publications. Testing program deliverables have been multiplying, while the testing cycle has remained essentially the same.

The investment by governments in standards-based assessment programs in Australia represents a significant component of central education budgets. It is not surprising then that Ministers expect timely and accurate results, and expect testing to contribute to raised standards and improvements in student learning. These are indeed high expectations to meet on an annual basis.

Much has already been written in assessment literature about the expanded use of data derived from tests. Many experts have noted how the tests were originally introduced to monitor student performance, but they have now become associated with decisions about teacher competence and school effectiveness to name just a few outcomes (Masters, 2004; Earl, 2000, 2001; Resnick et. al 2003-2004; Zenisky et al 2003-2004). The consequences linked to our test scores are growing every year. This has significant implications for the validity and reliability of the data generated by our testing programs.
Significant research into the consequences of the effects of increased accountability is already taking place in other parts of the world. In a recent study of the alignment between the standards and selected tests in a representative sample of American states, Resnick and others reported that,

\[\text{As test-based accountability becomes more stringent...the pressure to focus on the test will become even greater.}\]

(Resnick et.al 2003-2004, 25)

As school leaders find themselves under increased scrutiny we should expect a similar level of attention to be paid to our tests in Australia: especially if principals are to be removed from schools with poor levels of performance over time, as occurred in Victoria recently.

Full cohort testing in the compulsory years is now common practice around Australia and there is a high level of acceptance in the general community, and in many schools. However, there are still pockets of resistance around the country. In Victoria, for example, some schools continue to boycott the tests each year, and in other systems, test materials may be vandalized as a record of protest against centralized assessment and data collection.

All states now test their students in Years 3, 5 and 7 in literacy (reading, writing and spelling) and numeracy to monitor individual student progress and for the purposes of national and system level reporting. In addition, ACT, WA and Tasmania test Year 9 students, and all systems have added local curriculum elements such as viewing, speaking, listening, and/or and other learning areas, such as science in WA and computer literacy testing in NSW.
Other year levels continue to be added, Years 6 and 10 in NSW, and the Random Sample Assessment Program continues in WA alongside the Monitoring Standards in Education (MSE 9) program in WA. In such a short period of time, Australia has come to accept standardized testing in year levels one would never have assumed practicable or even necessary just twenty years earlier.

**Reporting test data**

As noted earlier, data derived from testing programs are now used and published in Australia for a variety of purposes and audiences. Education bureaucracies have made increasing use of data to monitor and report on the performance of students, schools and to hold teachers and schools accountable (Masters 2004). Like school data is used to judge the efficacy of regional governance and programs and to account for expenditure and distribution of resources.

Parents and guardians receive valuable reports on the progress of their child relative to expected standards and to the progress of other children of the same age or school level.

Education ministers receive benchmark data and report on student performance to Parliament: our statistics are translated into system performance measures and operational Key Performance Indicators. Ministers in turn are held accountable to Finance and Treasury for public expenditure on system level testing programs and are expected to produce gains and account for any slippage in student performance each financial year.

The media use student achievement data to heap public praise or provoke embarrassment. Some papers long to construct league tables akin to those used in the UK on the grounds of freedom of information and the public interest. In the US test data is now published on websites to assist parents in the selection of a local school and effects increases and decreases in student numbers, and the corollary, federal funding.

---

Presented at the 9th Annual Assessment RoundTable: Assessing Assessment Conference, Sydney, New South Wales, 7-9 November 2004
Australian governments too are using test data as a policy lever to drive resource allocation, school improvement and curriculum reform. Never before has standardized test data in the compulsory years of schooling been connected to so many educational outcomes.

**Testing numbers**

When combined the testing programs around the country amount to an extraordinary amount of testing, most of which occurs on an annual basis. In 2003, we put approximately one million students in the compulsory years to a vast array of tests (see Tables 1 and 2 below).

**Table 1.** Approximate total number of students in each State/Territory test in 2003. (N.B: totals are approximates as systems tend to report ‘more than’ and combined totals in annual reports)

**STATE/TERRITORY**

**NSW**

<table>
<thead>
<tr>
<th>Test Description</th>
<th>No. of students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Years 3 and 5 Basic Skills Test (BST)</td>
<td>165,000</td>
</tr>
<tr>
<td>Years 3 and 5 Primary Writing Assessment (PWA)</td>
<td>163,000</td>
</tr>
<tr>
<td>Year 6 Computer Skills Assessment (CSA)</td>
<td>59,000</td>
</tr>
<tr>
<td>- pilot on-line</td>
<td>23,000</td>
</tr>
<tr>
<td>Years 7 and 8 English Language and Literacy Assessment (ELLA)</td>
<td>141,000</td>
</tr>
<tr>
<td>Years 7 and 8 Secondary Numeracy Assessment Program (SNAP)</td>
<td>100,300</td>
</tr>
<tr>
<td><strong>TOTAL by tests</strong></td>
<td><strong>651,300</strong></td>
</tr>
<tr>
<td><strong>TOTAL students reported by year level</strong></td>
<td><strong>465,300</strong></td>
</tr>
</tbody>
</table>

**Victoria**

<table>
<thead>
<tr>
<th>Test Description</th>
<th>No. of students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Years 3 and 5 Achievement Improvement Monitor (AIM) test in literacy and numeracy</td>
<td>120,000</td>
</tr>
<tr>
<td>Year 7 Achievement and Improvement Monitor (AIM) test in</td>
<td>45,000* (based on</td>
</tr>
</tbody>
</table>
literacy and numeracy 2002 no.s)
Computer Adaptive testing (AIM on-line) 10,000*

TOTAL students reported by year level 175,000

Queensland

Aspects of Literacy Test Year 3
(Reading, Viewing, Spelling, Writing)
53,765

Aspects of Numeracy Test Year 3
53,765

Aspects of Literacy Test Year 5
(Reading, Viewing, Spelling, Writing)
54,172

Aspects of Numeracy Test Year 5
54,172

Aspects of Literacy Test Year 7
54,118

Aspects of Numeracy Test Year 7
54,188

Total by tests 324,110

Total students reported by year level 162,055

Western Australia

Western Australian Literacy and Numeracy Assessment (WALNA)

Year 3 (Reading, Writing, Spelling and Numeracy) 24,692
Year 5 (Reading, Writing, Spelling and Numeracy) 25,358
Year 7 (Reading, Writing, Spelling and Numeracy) 25,200

Total reported WA students 81,500

Tasmanian students and some independents schools from SA and ACT participating in WALNA

20,500

Total no of students reported as tested by WA tests 121,000

Random Sample Assessment, 10 % of Year, 3, 7 and 10 is tested in
2 of the 8 learning areas. (Approx. 10 % of cohort).
n/a

Monitoring Standards in Education Year 9 (MSE 9) 20,500

Total no of students reported as tested by WA tests 121,000
### South Australia

State Literacy and Numeracy Test (LaN Test)

Years 3, 5 and 7

Total students reported by year level **40,000**

### Australian Capital Territory

The ACT Assessment Program (ACTAP)

Year 3 literacy and numeracy (& viewing, listening and speaking)

Year 5 literacy and numeracy (& viewing, listening and speaking)

Year 7 literacy and numeracy

Year 9 Literacy and numeracy

Total no of students **18,700**

### Northern Territory

Year 3 literacy and numeracy

2,175

Year 5 literacy and numeracy

2,148

Year 7 literacy and numeracy

2,000

Total **6,323**
Table 2: Summary
Total number of students in 2003 testing program by State/Territory
(N.B: totals are approximates as many systems reported ‘more than’ and combined totals in annual reports)

<table>
<thead>
<tr>
<th>State/Territory</th>
<th>Number of Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSW</td>
<td>465,300</td>
</tr>
<tr>
<td>Vic</td>
<td>175,000</td>
</tr>
<tr>
<td>QLD</td>
<td>162,055</td>
</tr>
<tr>
<td>WA (incl TAS)</td>
<td>129,000</td>
</tr>
<tr>
<td>SA</td>
<td>39,000</td>
</tr>
<tr>
<td>NT</td>
<td>6,323</td>
</tr>
<tr>
<td>ACT</td>
<td>18,700</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>995,378</strong></td>
</tr>
</tbody>
</table>

Given the scope of compulsory years testing programs, the level of public investment and increased levels of accountability, all aspects of testing need to stand up to public scrutiny and should be subjected to stringent quality assurance.

Testing the quality of items: is there room for improvement?

*The success of standards-based education systems depend on two elements: strong standards, and assessments that measure what the standards expect.*

(Resnick et al, 1).

*For all, the high stakes for educators and students that will now be attached to test outcomes create new demands for technical quality and fairness in ...standards and assessment systems.* (Resnick et al, 3).
Given the volume and scope of testing outlined above, the number of items generated each year in Australia by individual States and Territories must be in the vicinity of tens of thousands.

Earl (2001) maintains that the value and utility associated with data come from the care with which the information has been collected and collated. Can the same be said for all test data in Australia?

Centrally set tests then, should model, rather than just measure, the standards expected for student learning. It is important not to lose sight of the initial purposes of the tests (to monitor student performance) and therefore we should ensure that our students remain central to our testing programs. All system tests should provide a fair measure, and be seen to provide a fair measure by stakeholders if we are to stand up to increased public scrutiny and accountability.

In a large scale study in the US of Differential Item Functioning (DIF) using data sets involving over 60,000 students, Zenisky, Hambleton and Robin produced several findings they considered to be worth reporting to test developers. The study involved grouping items by type, over time and levels, and identified attributes of items that can possibly contribute to differential item functioning. These include;

- The content
- Item type or format
- Item text
- Cognitive dimensions
- Negatively worded item stems; and
- The presence of pictures or other reference material.

As they found in the US, there are instances where unfairness is introduced into the test. As a result of the study, Zenisky and the other researchers concluded that linking of significant consequences to test scores, when the validity of test scores is questionable, is highly unsound. This begs the question of whether potential sources of DIF in our tests

Presented at the 9th Annual Assessment RoundTable: Assessing Assessment Conference, Sydney, New South Wales, 7-9 November 2004
are not just detected but are understood. How frequently do students produce incorrect responses as a result of poor test lay-out, confusion over negatively worded items and ambiguously worded instructions? Are there cases where the error is produced due to a language problem rather than a problem with computation? How do we report and evaluate these errors post administration? To answer these questions, I examined test items drawn from tests over the past four years from six State and Territory testing programs, a selection of these are provided for discussion at the session and have been anonymized in the interests of improvement rather than embarrassment. (See PowerPoint slides for sample items).

Improving student performance: some suggested ways forward

There is an expectation within the current testing context that we will continue to improve test scores. How might we achieve this? One important way is to improve the quality of the tests themselves, to improve fairness and hence the validity of the results. The hallmarks of quality test items are well known and documented and include:

- The item is closely aligned to the curriculum
- It involves clear and unambiguous instructions to candidates
- Use of appropriate level of language/literacy
- The items is clearly presented through choice of layout, visual cues, visual design, format and language
- The item tests what it is intended to test
- Absence of bias, cultural, gender or religious

When students have difficulty reading the question, but could have solved the problem or demonstrated the requisite learning, we have a poor test item and a validity problem.

One way to improve test items around the country would be for test developers to document best practice and to share it more systematically. There are a wide range of item development models being used with varying results around the country. The strengths and weaknesses of each could be shared for discussion at a future round table.
conference. A special interest discussion group and forum for developers to share their knowledge would be a positive addition to our current range of expert groups. Constant review and evaluation of existing practice is also essential for continuous improvement, even for those with many years of experience.

There are other models of good practice such as the publication of items with annotations to indicate the learning that students need to demonstrate when answering the question. These examples can also be helpful for training of item writers.

Similarly, the publication of assessment principles and guidelines for test development, and monitoring and evaluation of the achievement of those principles is an important component in quality assurance for systems. See the NSW K-6 website for an example, www.nsw.schools.ns.edu.au/learning/k-6assessments/principles.

Other strategies include publishing (as most systems do) reports on the previous year’s testing program and outlining the purposes of key stages; for example, reviewing and trialing of items. Publishing the characteristics of quality items is also beneficial for teachers and item writers. Another important quality assurance mechanism is a critical friend.

Further research into this area will assist with the improvement of student performance on standardized tests. Above all, it is essential that we maximize every child’s chance to demonstrate what they know, rather than what they don’t know. We need to aim high to produce fairer tests of student learning. The sample items selected for review indicate that there is definitely room for improvement and that above all, tests should provide a fair measure of the standard of learning expected for all Australian students.
REFERENCES


Department of Education and Training, *Western Australian Literacy and Numeracy Assessment 2003: Information to Teachers*


Presented at the 9th Annual Assessment RoundTable: Assessing Assessment Conference, Sydney, New South Wales, 7-9 November 2004
Putting the Nation to the Test


www.decs.sa.gov.au

www.eddept.wa.edu.au/mse/index.html

www.eddept.wa.edu.au/walna

www.education.tas.gov.au/annual report/03-04

www.qsa.qld.edu.au/testing/357tests/

www.schools.nsw.edu.au