

INTERNATIONAL TESTING & THE RESULTS OF AUSTRALIAN STUDENTS: Kevin Donnelly, May 19-20 *The Australian*: “Testing methods mask our failings”

Kevin Donnelly argues Australian students are performing in the second eleven in the TIMSS tests. Not only are we consistently outperformed by a handful of other countries, we have fewer students performing at the advanced level and a significant gap between strongest and weakest students. The main culprits are outcomes-based education and Australia’s national benchmarking.

THE FACTS - TIMSS

Firstly, this claim needs to be read in conjunction with the material on PISA (see next section). TIMSS does make Australia appear worse than PISA, so TIMSS is focused on by Donnelly and others who wish to construct a crisis in Australian education - with PISA being consequently disparaged.

Secondly, how does TIMSS actually get reported by ACER? TIMSS 2002 is reported by ACER (*TIMSS Australian Monograph #6*) as follows:

- a) Year 4 Maths – Australia above international average
- b) Year 8 Maths – Australia “significantly higher” than international average
- c) Year 4 Science – Australia “significantly higher” than international average
- d) Year 8 Science – Australia “significantly higher” than international average

This is not the language of crisis.

It is true that Australia has effectively stood still in these tests since 1994-5, while other countries have improved. Donnelly largely blames Outcomes-Based Education (OBE) on these results. Certainly, the standing still factor on the face of it is a cause for concern, but if a Research Methods 101 student found these results, they would immediately be sent to investigate:

- the comparative results between Australian states which have ‘strong’ OBE curricula and those who don’t. Donnelly doesn’t do that either in this report or in *Dumbing down*.
- What has happened in the countries that have passed Australia, such as a surge in investment in education? This is a central research question. Donnelly doesn’t ask it.
- Are there factors that differentiate performance, such as between urban and rural/remote areas? That is, are there SES factors in performance, such as the recent national benchmarking tests have shown? The answer is “Yes”, as comparative results between the ACT and the Northern Territory show – and, moreover, these are the same comparative results as are shown in PISA. Moreover, both Maths and Science results correlated positively with number of books in the home – another SES-linked result. Donnelly doesn’t even ask the question.
- Is there disinvestment in education since 1994-5 in countries like Australia that have stood still? Governments cannot disinvest in education, especially public education, and then blame others for weaker results – such as blaming OBE or teachers. Donnelly doesn’t ask that question.

TIMSS is not a cause for complacency, but neither is it a cause for crisis and any simple connections with OBE are meaningless.

Is there a difference between what TIMSS tests and what PISA tests? Yes – TIMSS tends to test the kind of content that can be learnt and memorised, while PISA tends to test for the application of knowledge: understanding, problem-solving etc. Australia tends to score comparatively better at the latter. There is no cause for complacency about the former, but neither is it a crisis and a student who blamed it on OBE would fail Research Methods 101.

THE FACTS - PISA

Kevin Donnelly disparages Australia's results in PISA, picking out of the ACER/OECD report only the statement that students' errors in spelling and punctuation are not accounted for in PISA.

As the reports from PISA in both 2000 and 2003 show, Australia is outperformed only by Finland in terms of statistical significance. To highlight only one statement from the ACER/OECD report distorts the findings. Australia's results are outstanding – that is the first and most salient point. The issue of not correcting faulty spelling etc is bizarre on two fronts:

a) PISA tests reading literacy – why spelling and punctuation should be accounted for in a test of reading comprehension is something Donnelly does not explain

b) Donnelly shows the most fundamental misunderstanding of assessment. Since PISA is a norm-referenced instrument; its results are comparative. If spelling and punctuation were to count as part of the scores in reading, they would have to do so for every country. In comparative tests, one condition cannot hold for some participants and not others. On the assumption that the spelling and punctuation factor would hold equally for every country – and there is absolutely no reason to believe otherwise – Australia would retain its comparative position.

National Literacy Testing: WHAT *Dumbing Down* SAYS

In both *Dumbing down* and his submission to the Senate Employment, Workplace Relations and Education Committee Inquiry Into the Standards of School Education, Kevin Donnelly argues that the 1996 national literacy test showed that approximately 27% of Year 3 students and 30% of Year 5 students failed to meet the minimum benchmark. Further, he states that research carried out by the Canberra-based academic, Andrew Leigh, after analysing the data that is available, concludes that standards are not as strong as they should be. Leigh states: "...troubling new evidence suggests that literacy and numeracy scores have stagnated or fallen since the 1970s – despite a doubling of resources"

THE FACTS

This study, conducted by ACER in 1996, was a national survey of the range of literacy achievements among Year 3 and Year 5 students in Australian schools. 4,000 students at each year level were surveyed.

The overall purpose of the survey was to provide baseline data to monitor national performance over time. Following the structure of the *National English curriculum profile*, the skills surveyed were reading and viewing, speaking and listening and writing.

Data for the survey was collected from an external assessor and a teacher assessing 10 students in selected classes over a six-week period on the basis of set tasks on all skills, combined with examples of best class work in writing and speaking. The results were then placed against "achievement levels" - there being 5 achievement levels in each of the areas writing, reading, viewing, listening, speaking. These achievement levels were constructed using the *National English curriculum profile* as a framework. The report's findings describe in detail what achievement consists of at each level. Year 3 and Year 5 students were assessed and reported on a common scale.

Results

- in each of Years 3 and 5, there was a large range between the highest and lowest achieving students: an estimated difference of five year levels between the top 10% and the bottom 10% in each grade. We will report on only reading and writing below.

Writing

- 5% of students in Year 5 were achieving at the highest level (5), ie very good control over grammar and punctuation in complex sentences, precise and effective vocabulary, coherent written arguments, engaging to read
- a further 33% of Year 5 students and 12% of Year 3 students achieved at Level 4, ie use both simple and complex sentences, use mostly appropriate punctuation, spell most words correctly, express a clear viewpoint and can construct a detailed storyline
- a further 35% of Year 3 students and 15% of Year 5 students were at Level 2, ie they write in a way which can be interpreted by others, shows some attempt at punctuation, but shows little shaping
- 6% of Year 3 students were at Level 1, ie they produce some recognisable words, but the writing is understood only by them at the time of writing

Reading

- 12% of Year 5 students read at the highest level, inferring meaning from figurative and idiomatic language
- a further 39% of Year 5 students and 12% of Year 3 students achieved Level 4, ie recognise tone, select relevant information, order events and recognise how linguistic features support implicit ideas
- a further 42% of Year 3 students and 21% of Year 5 students achieved Level 2, ie recognise main ideas, connect bits of information, predict a plausible ending
- at Level 1, 4% of Year 3 students were able to do little more than than predict the contents of a storybook from its cover and illustrations

In addition,

- girls outperformed boys in each aspect - most in writing and least in viewing
- NESB students averaged lower than English-background students
- parents' occupation was a highly significant predictor of literacy performance
- girls' and boys' achievement differences were greater in the lowest-coded occupational groups than in other groups
- Aboriginal students had average achievement levels 3 to 4 years below other Australian students
- Aboriginal students had 3 times the average rate of absence from school

Actually, two reports were released in 1997 based on this test. One was entitled *Mapping Literacy Achievement: Results of the 1996 National School English Literacy Survey*, and one was called *Literacy Standards in Australia*. The first was the full report of the ACER survey that had been in progress for the previous four years.

At the time of the release of *Mapping Literacy Achievement*, Australia was also involved in the beginning stages of mapping a series of "Benchmarks" which could be regarded as minimum satisfactory standards in literacy. As part of the study *Mapping Literacy Achievement*, ACER was asked to compare their results against the Benchmarks as they were then being developed. ACER did this for reading and writing only. They discovered that the then current benchmark definitions corresponded to particular ranges of achievement in the Levels they had defined. This overlaying exercise yielded the following results:

WRITING: 6% of Year 3 students below the benchmark
15% of Year 5 students below the benchmark

READING: 4% of Year 3 students below the benchmark
21% of Year 5 students below the benchmark

After this, however, in a second report, ACER was asked to establish one specific cut-off point ("cut score") for benchmark achievement. This cut-score was located above MOST tasks judged to be "below benchmark", but below MOST tasks judged to exemplify "benchmark performance or better". These results were published in the second document, *Literacy Standards in Australia*. The results yielded by this method of comparing the ACER findings with the benchmarks were as follows:

WRITING: 28% of Year 3 students below the performance standard (no longer 6%)
33% of Year 5 students below the performance standard (no longer 15%)

READING: 27% of Year 3 students below the performance standard (not 4%)
29% of Year 5 students below the performance standard (not 21%)

It was the second of these studies that the then Minister, Dr Kemp, took onto *60 Minutes* at the time to claim a crisis in Australian education. It is also the second of these results that commentators like Donnelly have used ever since to continue to construct the crisis.

ASSESSMENT: WHAT *Dumbing Down* SAYS

On pp. 41-42 of *Dumbing down*, Donnelly presents an argument on assessment that runs two sets of equivalences set out as a table, summarised as:

<i>Traditional Assessment</i>	<i>Outcomes-Based Education (OBE) Assessment</i>
norm referenced	criterion-referenced
summative	formative

THE FACTS

Donnelly shows a strong lack of understanding of some very basic principles of assessment here. He doesn't appear to understand that the norm-referenced– criterion-referenced binary and the formative-summative binary exist on different planes– they are concerned with entirely different things and thus cannot be equated in any way. The norm-referenced– criterion-referenced distinction is a distinction about *methods* of assessment : students ranked against each other or assessed against a set of criteria. The formative-summative distinction is a distinction about the *purposes* of assessment – for on-going diagnostic purposes *during* instruction (formative) or for final end-of-teaching-unit evaluation (summative). Hence, criterion-referenced assessment, for example, could be used for formative purposes or for summative purposes. In usual practice, *both* 'norm-referenced' and 'criterion-referenced' tend to be phrases used about *summative* assessment only, though it would be conceivable to have formative examples of each. Regardless of this, the two sets of binaries have nothing to do with each other, but Donnelly appears not to understand this. The 'traditional – norm referenced-summative' / 'OBE- criterion-referenced – formative' set of links is, simply, wrong. When on p. 187 of *Dumbing down*, he describes teachers 'checking their work, giving them feedback about strengths and weaknesses and modifying teaching to suit the needs of particular students' as 'an essential part of the traditional classroom', he is actually describing *formative* assessment, yet appears to believe he is describing *summative* assessment, because he is describing the 'traditional' curriculum.

NORM & CRITERION REFERENCING: WHAT *Dumbing Down* SAYS

Donnelly argues that in terms of assessment, norm referenced allows students to fail (because it is 'traditional'), while criterion-referenced allows all students to pass.

THE FACTS

Donnelly here once again shows the most elementary misunderstanding of what the terms 'norm-referenced' and 'criterion-referenced' mean in practice. In fact, the very opposite is the case. NSW moved its HSC into a variation of criterion-referenced assessment for this very reason. Norm-referencing is simply about ranking students in order. If a pass mark were set at, say, 50%, in theory, every student could fail by scoring lower than this pass. Yet under Donnelly's favoured norm-referencing, there would still be a ranking system; someone would still come first and someone come last and there would still be a bell curve, but all marks would be below 50. Someone has come first, but no one has passed. Similarly, if all students scored above 50, the scenario would be equivalent with no fails, but a student still in last position. In fact, norm-referenced assessment tells parents nothing about what their children know or can do – just where they stand in relation to other students. Whether this information is of any use to parents depends entirely on what information the parents have about how good or poor the *other* students are. Criterion-referenced assessment does not suffer from this problem. Contrary to what Donnelly appears to believe, only criterion-referenced assessment can make a statement about passing or failing if these words have anything to do with what students know or can do. One sets the criteria for passing : 'X amount of knowledge about Y' or "Ability do X'. Students either meet the criteria or they do not. Where they fall in relation to other students is irrelevant, but passing or failing is actually the point – and if all students do meet the

criteria, why should any of them 'fail' anyway? Similarly, if no students meet the criteria, why would any pass just to construct a bell curve?