Latest Developments and Trends in Assessment: Australian & International Perspectives

Gordon Stanley
Honorary Professor of Education, The University of Sydney
Emeritus Professor of Psychology, The University of Melbourne
Global reform trend

International tests, especially PISA drive reform akin to an Olympics quest for gold medals.

Globally how we collect data about student learning and achievement is at the centre of education reform.

Are our practices fit for the national purpose of human development?

Reporting required to be based on credible evidence about developmental progress and achievement.

Reported with reference to a series of standards
Role of standardized testing

International tests are usually sample-based and voluntary in terms of participation.

National tests are usually required as a condition of funding or registration.

Standardized exams can be used at school level as an outsourced or commercially provided service.

Students and their teachers, as well as tutors and coaching colleges, all take an intense interest in the process and outcomes.
What are the problems with standardized testing and the uses made of it?

The results of standardized testing are commonly used as a performance indicator of a school or education system rather than as a tool to assist student learning.

When standardized testing is used in high stakes situations few users of the results question or understand the limitations of the data obtained.
21st century: Integration of curriculum, pedagogy and assessment

Curriculum: organized around growth continua so that teaching and learning plans at each level build on past learning and direct current learning towards higher levels.

Pedagogy: classroom activities designed to assist student progression on growth continua, delivering higher outcomes for all.

Assessment: is a normal part of tracking student growth and providing feedback to assist further progress.
Curriculum described in terms of growth

Modern curriculum is organised around the idea of a developmental continuum.

Different growth continua attempt to define what it means to make progress or to improve in an area of learning or domain of knowledge.

Various locations on the continua are described in terms of what students ‘know and can do’ at that point (achievement standards).
What does growth mean?

Developmental continuum (or attribute structure) for showing growth in an area of learning

<table>
<thead>
<tr>
<th>Description</th>
<th>Level 1</th>
<th>Description</th>
<th>Level 2</th>
<th>Description</th>
<th>Level 3</th>
</tr>
</thead>
</table>

This presentation (C) Gordon Stanley 2016
Other terms used to describe growth

These are often referred to as:
- progress maps
- growth paths
- progress variables
- developmental continua
- progressions of developing competence
- profile strands
- increasing depth of knowledge
Good assessment practice

- **Assessment tasks** need to be designed to elicit student performance and enable it to be judged with respect to the intended outcomes.

- The syllabus and teaching program needs to be explicit about the *developmental continuum or progress map* in the area or field of study.

- The assessment program has to ensure that higher scores or higher grades are based on evidence about progress on the *developmental continuum*. 
Stress effects of standardized testing
‘PISA seeks to assess how well 15-year-olds are prepared for life’s challenges. ... focusing on young people’s ability to use their knowledge and skills to meet real-life challenges, rather than merely on the extent to which they have mastered a specific school curriculum.’ (OECD p.20)

PISA’s statement of aim indicates that the link between achievement and curricula is not regarded as the main objective of the study. PISA adopts a “literacy” concept about the extent to which students can apply knowledge and skills.
OECD statement on PISA

‘In a global economy, the yardstick for success is no longer improvement by national standards alone, but how education systems perform internationally. The OECD has taken up that challenge by developing PISA, the Programme for International Student Assessment, which evaluates the quality, equity and efficiency of school systems in some 70 countries that, together, make up 90 percent of the world economy.’
On the one hand OECD/PISA state that they do not measure according to school curricula, and not even the knowledge acquired at school.

On the other hand they claim that they do evaluate “the quality, equity and efficiency of school systems.”

It is also interesting to note that the importance of PISA is defined in terms of the fraction (90 percent) of the world economy, not in terms of the fraction of the world’s population.
Although PISA states that it does not test school knowledge, and that it does not test according to national curricula or testing school knowledge, the PISA results are interpreted, also in OECD reports, as valid measures of the quality of national schools systems, and the PISA reports are full of policy recommendations regarding schools (Loveless, 2009).

The PISA test is a pen-and-paper test, where students sit for 2 1/2 hours to answer written questions, in solitude and without access to sources of information.

How “real life” is this test situation?

How does it relate to the challenges that young people may face in their future life as citizens, as participants in tomorrow’s democracy and as skilled workforce?

Put in this form, the questions are rhetorical: the test situation does not resemble any real life situations.

At the country level, one factor is the “small - n”- problem: Given the low number of units of investigation and the large number of potentially relevant independent variables, it is not possible to isolate the effects of each independent variable on the dependent variable. From a statistical point of view, it is therefore not possible to answer the question why some countries perform better than others in terms of achievement scores. (Haahr, 2005, p 21).
PISA is by design culturally biased and methodologically constrained to a degree which prohibits accurate representations of what actually is achieved in and by schools. Nor is there any proof that what it covers is a valid conceptualisation of what every student should know.

The product of most public value, the national league tables are based on many weak links.

League tables depend on assumptions about their validity and reliability which are unattainable.
PISA: Australian data

Australian PISA scores, 2000-2012

Reading has been tested since 2000. Maths testing was introduced in 2003 and science testing was introduced in 2006.
TIMSS: Australian data

Trends in International Mathematics and Science Study (TIMSS), 1995-2011.

Data from the Trends in International Mathematics and Science Study (TIMSS) show that achievement in both mathematics and science at Year 4 and Year 8 has been stable over the past 16 years.

Source: ACER. Get the data
NAPLAN

• The Australian National Assessment Program – Literacy and Numeracy (NAPLAN) involves a full cohort testing of students in years 3, 5, 7 and 9 on reading, writing, language conventions and numeracy.

• The program was approved by State Ministers on the assumption that the tests would provide diagnostic feedback to schools, teachers, students and parents to assist in improvement in teaching and learning.

• School results are publicly available on the MySchool website where like schools test outcomes can be compared.
What is tested, and how?

- NAPLAN tests the sorts of skills that are essential for every child to progress through school and life, such as reading, writing, spelling, grammar and numeracy.

- The content of each test is informed by the *National Statements of Learning* for English and mathematics which underpin state and territory learning frameworks.

- Questions are multiple-choice or require a short written response. The Writing task requires students to write a narrative or persuasive text.
NAPLAN assessments are undertaken by students as a normal part of their education program.

Teachers will ensure students are prepared for the tests and will provide appropriate support and guidance.

The use of services by coaching providers is not recommended.

The most effective way to prepare for the tests is to ensure that literacy and numeracy skills embedded in the normal curriculum are taught.

Sample questions and a sample Writing task can be accessed on the NAP website.
NAPLAN coaching texts
But ACARA chief executive Robert Randall said there was no evidence that "excessive coaching or excessive drilling" had improved results, and teachers and students should treat NAPLAN testing days like any other day on the school calendar, like a swimming carnival or cross-country race.

"Drilling and excessive practice around sample tests aren't helpful," Mr Randall said.

"Once you've done a couple of practice ones to familiarise the students, you are losing any gain that you get from it. It becomes a distraction and we think it's unhelpful.”

Source: SMH 9 May 2014
NAPLAN data show that there was little change in student achievement between 2008 – 2016.
Pattern with means

For each grade level and each domain shown no evidence of consistent improvement (upward trend) from 2008-2016.

Essentially flat profile over time with small fluctuations.
Year to year trend effect size

- For each domain and grade level the difference in mean from the preceding year as a proportion of sd is calculated.
- This measure is called an effect size.
- If each new year grade 3 was improving relative to last year grade 3 the effect size for grade 3 would be trending upwards over the years.
- A similar pattern would be expected for each grade level over time.
Pattern with year to year trend

- Effect sizes fluctuate ranging from positive to negative shifts for a grade level over time. This may be caused by variation in equating error over time swamping out any growth trend.

- At an aggregate total population level one would not expect much true score fluctuation over time.

- However if true scores were improving through better teaching/learning then upward trend should be detectable.
Equating tests

• Equating one test with another is a complex technical procedure and involves some degree of statistical error.

• There can be minor fluctuations in the average standardized test results from year to year when, in reality, the level of student achievement has remained essentially the same.

• It is only when there is a meaningful change in the results from one year to the next, or when there is a consistent trend over several years, that statements about improvement or decline in levels of achievement can be made confidently.
What can we make of this?

After 8 years of NAPLAN testing it would seem that results comparing the total population from year to year indicate no clear trend of growth in outcomes.

Given the range of error which can operate in large scale testing programs one must ask whether NAPLAN tests are sensitive enough to measure any real changes in performance.
Margaret Wu has canvassed the issues of the problem with NAPLAN tests not being sufficiently reliable to act as an accountability measure of teaching.

ICT in assessment systems

Information and communication technology (ICT) offers promise of greater process reliability and opportunity for innovative assessment.

High stakes systems need to be risk averse, so progress in integration of ICT has tended to be evolutionary rather than revolutionary.

Examination systems are now confronted by students who are digital natives and who are disadvantaged by methods that rely on pen and paper inputs.
ICT based assessment

Many ICT based tools and systems are becoming available.

Some are transitional in using ICT to augment existing testing and assessment.

Due to easy scalability with ICT much interest in developing integrated teaching/learning/assessment packages.

Possibility of real time personalized assessment for learning.
Assessment using ICT

- International tests piloting on-line delivery.
- National tests going on-line:
  - RTT
  - NAPLAN
- Commercial on-line teaching/learning/assessment
ICT provides more options

Simulations

Richer test items: incorporating video
different response formats

e-Portfolios

Gaming formats

Machine processing of results and personalized feedback.

Automated essay scoring (AES)

Conversation-based assessment (CBA) systems
21st Century assessment

Technology does not replicate paper exams but allows new, more valid types of assessment.

Skills that cannot be assessed adequately without performance are captured ‘in action’, using technology where possible.

Personalized and adaptive assessment which confirms skill level.
Automated essay marking systems (AES)

When students keyboard rather than handwrite the viability of automated essay marking becomes real.

There are hundreds of computer tools and algorithms that quickly analyse texts and large text corpora.

Automated essay scoring has reached an accuracy level such that the scoring of many classes of written essays is as accurate as the scoring of human raters.

Corpus dependent systems (e.g. LSA) require training on exemplar text.

Corpus independent systems draw on large scale corpora.
Automated essay marking systems (AES)

Critics voice concerns about differences in the criteria that humans versus computers use to grade the essays.

The incremental gain from advanced computational algorithms beyond word count is sometimes unspectacular.
Criticism of AES

Shermis e.g:

So here we have the discovery of America: Queen America sailed to Santa Maria with 1492 ships. Her husband, King Columbus, looked to the Indian explorer, Nina Pinta, to find vast wealth on the beaches of Isabella, but would settle for spices from the continent of Ferdinand.

According to Shermis most current systems would grade that as pretty much on target because it has ‘the key words and synonyms, and has good structure there’

From what we know of memory processes how likely is a student to produce this ‘word salad’ like answer?
AutoTutor

Art Graesser at the university of Memphis.

A fully automated computer tutor that assists students learning about hardware, operating systems, and the Internet in an introductory computer literacy course.

LSA is a major component in evaluating the quality of student contributions in the tutorial dialog.
How is the packet switching model of message transmission like the postal system?

Packets are sent to intermediate destinations before being routed to their final destination.
Table 2: Amount of Fiber in Food

<table>
<thead>
<tr>
<th>Food</th>
<th>Grams of Fiber</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 cup of olives</td>
<td>1</td>
</tr>
<tr>
<td>1 small apple</td>
<td>3</td>
</tr>
<tr>
<td>1 ounce of almonds</td>
<td>3</td>
</tr>
<tr>
<td>4 ounces of sweet potato</td>
<td>4</td>
</tr>
<tr>
<td>1 cup of chopped carrots</td>
<td>4</td>
</tr>
<tr>
<td>16 ounces of vegetable juice</td>
<td>4</td>
</tr>
<tr>
<td>2 slices of rye bread</td>
<td>7</td>
</tr>
<tr>
<td>2 cups of green beans</td>
<td>8</td>
</tr>
<tr>
<td>1 cup of raisins</td>
<td>10</td>
</tr>
<tr>
<td>1 medium avocado</td>
<td>11</td>
</tr>
<tr>
<td>4 cups of broccoli</td>
<td>12</td>
</tr>
<tr>
<td>1 cup of cooked lentils</td>
<td>16</td>
</tr>
<tr>
<td>2 cups of cereal</td>
<td>20</td>
</tr>
</tbody>
</table>

The ratio of the number of cups of raisins to the number of grams of fiber is ___________

1 to 10
Thank you for listening

Email: gordon.stanley@sydney.edu.au