Leading education and social research
Institute of Education
University of London

IAEA Annual Conference 2010

Bangkok, Thailand

Assessment fit-for-the-future

Professor Gordon Stobart

Emeritus Professor of Education

Institute of Education, University of London. UK

Professor of Education, University of Bristol, UK

g.stobart@ioe.ac.uk

The theme of this conference is 'Assessment for the Future Generations'. Implicit in this theme are ideas about meeting the needs of our current students, who will be the next generation, and those of those who follow them.

Much of our contemporary debate about preparing for the future is about how we develop and use our resources in a way that will leave a constructive legacy for future generations. This focuses on the *sustainability* of the earth's resources and on finding alternative technologies to prevent the further depletion of some natural resources. How do we reduce global warming so that future generations are not faced with environmental crises around water and extreme conditions? What can we do to reduce our dependence on oil?

In this context what might *sustainable assessment* look like – assessment that constructively prepares students for the future, rather than simply repeating wasteful patterns from the past? David Boud has described sustainable assessment in terms of:

> Any assessment act must contribute in some way to learning beyond the immediate task...assessment that meets the needs of the present and prepares students to meet their own future needs. (2000, pp.8-9)

Boud has developed this argument through his valuable concept of the *double duty* of assessment, in which assessment activities 'have to focus on the immediate task and on implications for equipping students for lifelong learning in an unknown future...they must attend to both the process and substantive content domain' (2002, p.9). Simply assessing for the here-and-now (assessment as a 'snapshot') is insufficient if nothing is carried forward, so too is process-based learning (for example, 'learning to learn', 'critical thinking') if there is no substantive learning in the here-and-now.

My intention in this paper is to develop this line of argument in relation to our current assessment practices so that we can begin to develop more sustainable ways of assessment, ways which not only offer a dependable assessment of current learning but also help students develop their own assessment resources. I focus on external assessments such as public examinations and national tests, but a fuller treatment of this theme would also consider how might also develop sustainable informal classroom assessment.

2010 – Where have we got to with testing?

If we review progress in testing and examinations[1] it could be argued that little has changed since the civil service selection examinations developed by the Ming Dynasty over five hundred years ago – which in turn had built on an examinations tradition developed over the previous five hundred years. So sitting down in an examination room, receiving the same question papers on specified subjects, being allowed the same length of time to complete the test and then being marked by unknown markers is nothing new. Indeed we seem to have got less demanding – in the Chinese examinations the candidate would be locked in a cell for several days (to prevent cheating) and the scripts copied by someone else (so that handwriting would not be recognised) and double marked. Like today, because the exams were high-stakes as they led to selection for the imperial civil-service, there were ingenious forms of cheating including miniaturised books of answers (small enough to get through the required body search) and printing answers on the lining of clothes.

What has changed is the scale and accessibility of such tests. Written assessments are now a world-wide phenomenon, taken annually by tens of millions of students. This has been possible because of the gradual development of testing systems and organisations which can handle the volume of data produced. The dilemmas that come with this are essentially about the validity of the tests themselves (see below).

The lesson to be drawn from the examination tradition is that assessment systems change incrementally rather than radically. There are radical developments, for instance in the area of computerised adaptive testing (see Bennett,1998), but their application tends to be specialist rather than mainstream. Similarly there have been alternative forms of assessment introduced, for example, portfolios which again have had only limited implementation.

The quality and impact of current tests

Before we ask what we can do for future generations, we need to check what we are doing for the present one. What will they carry forward to their unknown future from the assessment regimes that are shaping their identities as learners? The importance of external tests for progress within, and beyond, education has meant that assessment is central to learner identity – we are shaped and defined by our assessment results. As Allan Hansen (1994) has tellingly put it 'The individual in contemporary society is not so much described by tests as constructed by them'.

[1] I use the two interchangeably – with tests as shorthand for both open-ended written examinations and fixed response computerised tests.

This is not a call to abolish tests and examinations, simply to try to make sure that there impact is as constructive as possible in terms of what is taught in preparation for the assessment and what students carry away from the course that is examined.

What are the threats to achieving this in our current assessments? My own risk-list no doubt reflects my own experience of British-style examinations and of US style machine marked fixed response tests – but may also generalise to other contexts.

1. *They are of limited dependability*.
2. *They encourage a past-paper tradition of practice and recall.*
3. *The grades are more important than the knowledge tested.*
4. *Too much is read into the results.*

Test Dependability

Any test involves a trade-off of construct validity (the domain or skill being tested), reliability and manageability. Dependability is the optimal trade-off of these. In large scale high-stakes testing this trade-off often puts great emphasis on reliability and manageability - at the expense of construct validity. So a language examination is reduced to reading and writing in the language, neglecting speaking and listening – which may be central to the construct of learning a language. In this way the examination suffers from what Messick (1989) has called 'construct underrepresentation'. Dependability suffers because the test scores are not an accurate representation of someone's performance in that domain or skill – since key elements went unassessed.

This tension can be illustrated through the one-handed clock (Fig.1 – See Stobart, 2008, pp. 110-11 for fuller treatment). Where the hand is placed will always be at a cost to one of elements – construct validity, reliability or manageability. So a task could have a high construct validity – a demonstration of the skill in question – but of low reliability because there was no agreed scheme of assessment. Similarly, a multiple-choice test could be highly reliable in terms of marking, but have little or no construct validity if it was assessing reflective writing. In both cases there is low dependability.

What we are looking for is a trade-off in which the construct validity of the task is not over-compromised by reliability and vice versa. Manageability enters into this too, I can produce highly dependable assessments that would be too expensive to operate, for example requiring an external assessor to evaluate each student.

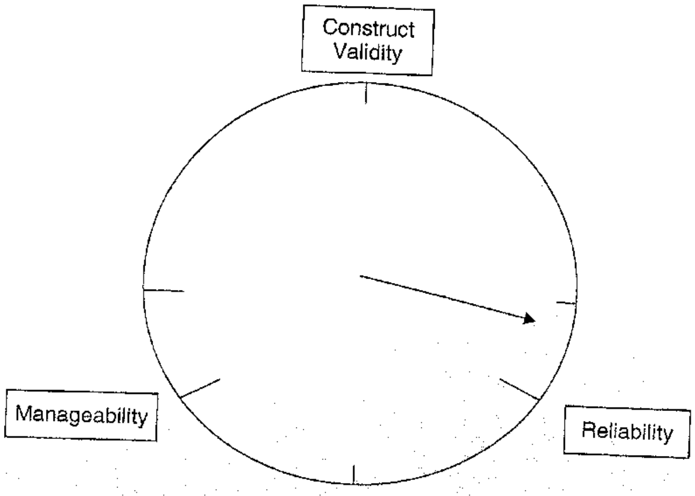

Figure 5.1 The one-handed clock.

Where do we want the hand to point? High construct validity and reliability (0–20 minutes) may cost in terms of manageability, while validity and manageability may be at a cost to reliability (40–60 minutes). What we should be attempting to do is to keep assessment out of the reliable and manageable zone (20–40 minutes) where it is often found. This is because it is often a very weak signifier of the skills that we want to assess, and it has limited construct validity, but it is chosen because it is both cheap ('efficient') and reliable. The 20–40-minute zone is most likely to produce negative backwash effects.

Different purposes may lead to different hand positions. I want my pilot at 'ten minutes past', regardless of expense, that is, with both realistic simulation training and actual flying, plus rigorous assessment of these. I will settle for '20 minutes past' for a national maths test, although this likely to miss out on applied skills. In English I may want to incorporate high validity and manageable elements such as teacher assessed speaking and listening which may weaken reliability.

So we will return to the question of how we improve the dependability of assessments for the future.

The past paper tradition of practice and recall

If an assessment is highly predictable from year to year it encourages teaching practices which focus on past papers and practising answers to previous papers. This encourages looking for cues and responding in a practised way. The risk is that this produces a surface learning that hardly lasts beyond the test. The threat to validity here is that of Messick's 'construct irrelevant variance' – the tests are measuring something other than what they claim to measure. Garrison Keillor offered a wry example in his *Lake Wobegon Days*:

> For years, students of the senior class were required to read ["Phileopolis"] and answer questions about its meaning etc. Teachers were not required to do so, but simply marked according to the correct answers supplied by Miss Quoist, including: (1) To extend the benefits of civilization and religion to all peoples, (2) No, (3) Plato, and (4) A wilderness cannot satisfy the hunger for beauty and learning, once awakened. The test was the same from year to year, and once the seniors found the answers and passed them to the juniors, nobody read "Phileopolis" anymore.

A more telling example was offered by Gordon and Reese from their research on how teachers and students prepared for the high-stakes *Texas Assessment of Academic Skills.* They found that direct teaching to pass the tests can be very effective, so much so that students could pass tests:

> …even though the students have never learned the concepts on which they are being tested. As teachers become more adept at this process, they can even teach students to answer correctly test items intended to measure students' ability to apply, or synthesise, even though the students have not developed application, analysis or synthesis skills (1997, p.364)

It would not be difficult to produce examples like these from assessment systems all round the world – indeed teachers pride themselves on being to spot what might come up in an examination and help their students by rehearsing prepared answers. This is also one of the most difficult elements to change in an assessment system, too much difference from previous years will bring both public and school outcries about questions being unfair ('we had not prepared our students for this').

Can we change this for future generations?

Grades rather than knowledge

What accompanies the past paper tradition is usually an emphasis on grades (or levels or percentiles etc.). This is inevitable when the main purpose of assessments is to select and/or become an accountability measure. The students want to optimise their grades and the schools want to impress with their results. The problem with this is that, in terms of assessment's 'double duty' very little, in terms of knowledge and skills, is carried forward. This is especially so when the assessment has limited validity and is a weak representation of the intended skills. This is what Allan Hanson calls the *fabricating quality of tests.* He claims that:

> The fabricating process works according to what may be called the priority of potential over performance. Because tests act as gatekeepers to many educational and training programs…the likelihood that someone will be able to do something, as determined by the tests, becomes more important than one's actually doing it. People are allowed to enter these programs, occupations and activities only if they first demonstrate sufficient potential as measured by tests. (p.288)

We will not remove the need for grades, what sustainable assessment may have to do is make the grades better signifiers of what knowledge and skills have been learned.

Interpreting the results

Current theorising of validity places the emphasis on the inferences drawn from the results. If the results of a well constructed and reliably marked test are used wrongly or misinterpreted, then the assessment is not valid. My extreme example is that if results on a well constructed maths test are used as the sole criterion for selection to Art School, we would immediately say this is not a valid selection tool. This is because you cannot infer from a maths score someone's artistic ability.

My claim is that many of our formal assessments suffer from *over-simplistic interpretations.* This is because in many education systems they are used as indicators of school achievement. This is particularly the case when test results are equated with educational standards – as they have been in England. So an improvement in national test results is interpreted as an improvement in national educational standards in that subject, even though other measures may suggest standards have not improved as much as claimed (Tymms, 2004).

Such interpretations in turn encourage schools to maximise their results. This may encourage harder work and higher expectations of students, but it can soon also lead to some fairly cynical 'playing the system' (see Stobart, 2008 ch.6 for a fuller account). This may include manipulating which students are entered and choosing options that will maximise results, irrespective of their educational value. Such strategies, along with constant practice in test-taking, then generate *score inflation,* with governments claiming credit for dramatic improvements in educational standards.

So how do we move away from simplistic interpretations of results so that individuals and schools might be better evaluated?

Strengthening assessment's 'double duty'

The review of where we are now has highlighted some of my concerns about current testing systems. I have raised some questions for the future about each of these. The task here is to explore some of the ways in which these might be answered in order to increase their validity in the here-and-now while leaving the test takers better prepared for 'an unknown future'.

How can we improve test dependability?[2]

1. *Make explicit the purpose and learning demand.* Achievement tests are assessments of something, typically a curriculum or subject or skill specification. A key validity question is 'what is the principal purpose of this assessment? To answer this we have to look at both the aims of the assessment and how the results are used. *It is the aims of the course, rather than its content, that should determine the purpose and form of its assessment*. The educational philosopher John White has shown how those developing the subject specific national curriculum programmes of study in England paid little attention to the declared aims and values of the curriculum, which had only been developed *retrospectively* after the first version. So, while the aims are about fostering curiosity and collaborative working, the programmes of study are overwhelmingly about content. This weakens coherence, which is then further undermined by even more restricted assessment - in which, for example, the applied elements of mathematics and science, and the speaking and listening in English are not tested.

   Learning demand reflects this concern with the broader aims of the assessment. What level of knowledge or skill matches these intentions? Some assessment systems use Bloom's Taxonomy as a hierarchy of cognitive demand, with its movement from knowledge through comprehension, application, analysis, and synthesis to evaluation. This has been challenged but generally serves as a useful framework, and analysis of tests using this will often identify how many questions are at the lowest level, recalling knowledge rather than showing an understanding of it. Bigg's SOLO taxonomy would be another productive possibility. This first practical step is about how the assessment meets the learning intentions or aims of the course/curriculum, rather than about the content coverage which so often dominates test construction.

[2] This section draws directly on Stobart, 2008, pp.105-115

2. *Encourage 'principled' knowledge through less predictable questions.* The suggestion that an examination should include unfamiliar questions or material, so that students have to rely on their understanding in order to fashion an answer, may sound innocuous but in practice will meet concerted resistance. In any past paper tradition, the reliance on predictability runs deep. Much of the preparation is about 'when you see this…', which shifts the emphasis to cue spotting and recall of prepared answers.

   My aim is to move from '*when you* …' to '*what if* …?' preparation which encourages more active student involvement and encourages a problem-solving outlook which may better serve students in their unknown future. One practical approach might be to gradually introduce less predictable questions in areas which have been announced in advance, indicating that the format may be one of several possibilities. I realise that this is already done in many syllabuses – but I also know that it is often not acted on in practice.

   If we consider *teachers' classroom assessments*, these may provide rich opportunities for developing more flexible and active forms of learning. The teacher knows what has been taught, though not necessarily what has been learned, and can therefore devise questions to see how well the learners are able to use it in an unfamiliar form or context. This not only tests 'principled' knowledge (which can be transferred to new situations) it provides feedback on misconceptions that would not necessarily be revealed by more predictable 'recall' answers.

   An important principle here is that *classroom-based tests do not have to continuously mimic the external tests* – which are always likely to be more restricted*.* So the hand can move nearer 12. While students will have needed to practise the particular examinations skills and formats, they do not need to practise these on every test for the whole course. However teachers are unlikely to change their practices unless there is some encouragement from changes in the external examinations.

3. *Keep it as authentic as possible.* If we want the backwash from tests to lead to teaching and learning practices which help develop the intended skills, then the more directly a test assesses these skills, the more likely it is to encourage them. Our aim has to be to keep the one hand (see Fig. 1) as near to 12 as possible.

The task is therefore to produce what Frederiksen and Collins have called a *systemically valid test*:

> One that induces in the education system curricular and instructional changes that foster the development of the cognitive skills that the test is designed to measure. (1989, p.27)

This is a good test of any assessment and an essential part of sustainable assessment.

This also has implications for what can be inferred from the grades awarded. Grades are critical in most selection and accountability processes. However if our assessments bear little relation to the real-life skills and understanding they are supposed to be measuring, then grades will be misleading and of limited validity. The more authentic the assessments are, the more confident we can be that a good grade represents good skills and understanding.

4. Make more intelligent use of assessment data.

Our assessment systems generate vast amounts of data on students' performances, little of which is utilised. In England, for example, the 11 year old cohort takes national tests, the principal use of which is to provide an average score for each school of the pupils reaching a given level ('80% at level 4 and above'). [The tests are not used for selection to secondary schools as they come too late for that]. This percentage is the key indicator in school accountability.

All this encourages 'playing the system' and extensive test preparation for the Year 6 (11 year olds) taking the test. What is required is a move towards more 'intelligent accountability' (O'Neil, 2002). In relation to the way assessment data are used I have proposed seven steps that could be taken:

i. *Set realistic targets.* Many governments and policy makers use test results to set targets. These are generally aspirational and unrealistic – the No Child Left Behind (NCLB) legislation in the US calls for all children to reach basic achievement level by 2014, even though a large proportion are not there currently. Robert Linn has proposed a model based on the principle that *performance goals should be ambitious, but should also be realistically obtainable with sufficient effort* (2005, p.3). He calls for an *existence proof*, evidence that the goal does not exceed one that has been achieved by the highest performing schools – so if they improved by 3 per cent each year over the last few years, that might be a realistic state goal. I would call these *empirical* targets.

ii. *Multiple measures.* To rely on a single measure, a 'headline' results summary, is inviting trouble, both in terms of how it will distort the system and of the consequences of its limited validity. What gets squeezed out in this are other measures of the quality of schooling such as teacher assessment, pupil satisfaction, absenteeism, and 'value added' measures of progress. Intelligent accountability would look for a more valid use of these data, joint reporting of teacher and test judgements or, better still, some reconciliation based on local discussion of the evidence.

iii. *Monitor national standards separately.* Using national test and examination results to determine changes in national standards is too much for them to bear. A more constructive way of monitoring national standards is to take a representative sample of pupils and use low-stakes assessments (individual and school scores are not reported because it is only a sample) which have common items from year to year. This reduces both preparation effects and makes comparisons between years more reliable. This is the logic behind the National Assessment of Educational Progress (NAEP) in America, the National Education Monitoring Programme (NEMP) in New Zealand, and the Scottish Survey of Achievement (SSA)

iv. *Include all – but not necessarily using the same measures.* In high stakes accountability systems based on test results there is the temptation to exclude those who are 'off the scale'. This can lead to neglect of vulnerable groups (NCLB, to its credit, deliberately includes these groups). What may be needed is more diversity of measures so that we have finer grained measurements against which to chart what may be much slower progress. Intelligent accountability would then monitor progress against this scale rather than an unhelpful report of little or no progress on a cruder scale. This may make reporting more complex, but with more sophisticated measures we expect that.

v. *Continuously evaluate the accountability system.* What an accountability system values will affect what goes on in schools. So it is important to review how accountability is modifying teaching and learning and any unintended consequences. It also means moving systematically away from narrow targets, which may narrow a school's focus, towards more sustainable changes in curriculum, teaching and learning which will be reflected in more complex and qualitative approaches to accountability. It also involves monitoring the reliability of the assessment system.

vi. *Monitor measurement error.* External assessments strive to be as reliable as possible, but there will always be measurement error which will result in misclassification. We need systematic checks on reliability. The more difficult issue is how much should be made public (see Newton, 2005) – something Ofqual, the regulator of examinations in England, is wrestling with at this moment.

vii. *Check unintended consequences.* This brings us back to where we started: any assessment has an impact, and the more high-stakes the assessment the greater the impact. If part of the purpose of an assessment was to raise standards and improve teaching and learning, is it having that effect or has it generated unwanted practices. This requires impartial evaluation of what an assessment policy has stimulated.

Conclusion

If we are to create sustainable assessments which do the 'double duty' of assessing in the here-and-now as well as generating lifelong knowledge and skills, we need to review our assessment systems. What are our assessments actually encouraging? This paper has looked for ways in which our assessments might encourage more flexible and active ('mindful') learning. This will include improving the dependability of our current assessments, with a move towards more authentic assessments and less dependence on practised answers to predictable questions. We also need to look for more intelligent accountability systems so that less depends on raw assessment results. If we want future generations to show the flexible skills that modern society demands, we have to encourage them in our assessments.

References

Bennett, R. E (1998) *Reinventing Assessment: Speculation on the future of large scale educational testing,* Princeton N.J. ETS.

Boud, D. (2000) 'Sustainable Assessment: Rethinking Assessment for the Learning Society', *Studies in Continuing Education*, 22: 151–167.

Boud, D. (2002) 'The Unexamined Life is Not the Life for Learning: Rethinking Assessment for Lifelong Learning', Professorial Lecture given at Trent Park, Middlesex.

Frederiksen, J. R. and Collins, A. (1989) 'A Systems Approach to Educational Testing', *Educational Researcher*, 18 (9): 27–32.

Gordon, S. and Reese, M. (1997) 'High Stakes Testing: Worth the Price?', *Journal of School Leadership*, 7: 345–368.

Hanson, F. A. (1994) *Testing Testing: Social Consequences of the Examined Life*, Berkeley, CA: University of California Press.

Keillor, G. (1985) *Lake Wobegon Days*, New York: Viking.

Linn, R. J. (2005) 'Issues in the Design of Accountability Systems', in J. L. Herman and E. H. Haertel (eds) *Uses and Misuses of Data from Educational Accountability and Improvement.* Chicago, IL: National Society for the Study of Education.

Messick, S. (1989) 'Validity', in R. L. Linn (ed.) *Educational Measurement*, 3rd edn, New York, NY: American Council on Education and Macmillan, pp. 13–103.

Newton, P. (2005) 'The Public Understanding of Measurement Inaccuracy', *British Educational Research Journal*, 31: 419–442.

O'Neill, O. (2002) *A Question of Trust*, Cambridge, UK: Cambridge University Press.

Stobart, G. (2008) *Testing Times, The Uses and Abuses of Assessment,* Abingdon, Routledge.

Tymms, P. (2004) 'Are Standards Rising in English Primary Schools?' *British Educational Research Journal*, 30 (4): 477–494.

White, J. (2004) *Rethinking the School Curriculum: Values, Aims and Purposes*, London: RoutledgeFalmer.