Evidence on the influence of assessment on teachers' classroom practices

Alan Bell and Hugh Burkhardt for the Toolkit team

Summary

How true is WYTIWYG? (*What You Test Is What You Get)* This tool brings together evidence of the effect of modes of assessment on teacher approaches in the classroom, and on student approaches to learning. In addition to reports on practice, two research studies are outlined. One is from a large-scale comparison of assessments and practices in two Australian states. The other is a study of the approaches to learning of students in higher education when they are assessed by essays, or by multiple choice tests.

The resemblance between the tasks in a system's high-stakes assessment and the focus of teaching and learning in most classrooms is not enough to show causal influence, either way. But the available evidence on introducing *changes* in such assessment suggests that the influence can be strong. In particular:

- When curriculum-embedded components are introduced into high-stakes assessment, they happen – teachers and students spend at least the expected time, working on the new component.
- When changes are made in external high-stakes assessment, introducing new task-types, this can have profound effects. The first research study reported, from Australia, showed that the introduction of non-routine problem solving into the Grade 11-12 assessment resulted in such work appearing in classrooms from Grade 8 upwards.

The evidence thus underlines the importance of the close alignment of any system's high-stakes assessment with its curriculum standards and learning goals.

The report also discusses the tendency of innovations to degrade in quality over time, the reasons for this and how it may be minimized.

Purpose

To outline review evidence of the influence of high-stakes assessment on the pattern of classroom practice in schools – and thus of the likely effects of choices of assessment by system leadership on the implemented curriculum in their schools.

Background and context

It is a widespread source of frustration to educators that classroom practice generally fails to match the aspirations of curriculum guidance. This is often attributed to the failure of assessment to demand and value the higher level, less traditional, mathematical activities. WYTIWYG, 'What You Test Is What You Get' is seen as a prime cause of this degeneration from the desirable. Even teachers who share the broader aims can feel constrained to 'do their best by their students' by coaching them narrowly for the examination.

In some systems professional development activities and/or the periodic visits of school inspectors provide an incentive to teach the full range of specified activities. Occasional inspections can not, in any case, match the persistent pressures for 'improved test scores'.

What evidence is there on this issue? There is some though, as usual, well-focused research studies are too few.

There are many reports from around the world that show a close similarity in range and balance between the set of tasks in high-stakes assessment and those that students face in their classroom learning. That is not enough to show any *causal* effect of assessment on the implemented curriculum.

The clearest evidence of such an effect will come when there are changes in the assessment system, without any simultaneous change in the *intended curriculum*, as specified in official documents. Then it is likely that any changes in the *implemented curriculum* in typical classrooms that resemble changes in the tests come from teachers' response to that change. Here we shall confine our review to such changes.

In looking at evidence, we shall distinguish between those changes that are:

- required as part of the assessment process itself – notably curriculum-embedded 'portfolio' elements that are produced by the student in the classroom as part of their work through the year;
- introduced into the pattern of classroom activities by teachers to help their students perform better in the new aspects of external high-stakes assessment.

Curriculum-embedded assessment components

There is overwhelming evidence that, when curriculum-embedded assessment is introduced as part of a high-stakes assessment system, it happens. The assessment innovations in the 1990s in US, notably Vermont and Kentucky, exemplified this. They showed key features of such a change – that teachers and students become part of the assessment process, in the choice of tasks for students to attempt and in evaluating their responses. (They also showed a feature of all assessment – degradation of the quality under a variety of pressures that are always there. We shall return to this below)

The British have more substantial experience of curriculum-embedded portfolio assessment (they call it 'coursework'). It began in the 1960s with the introduction of the CSE[1] examinations, aimed at students outside the highest-achieving quarter who took the "O-level"[2] examinations at Grade 10. The limitations of timed-written examinations in assessing the whole range of desirable aspects of performance were recognized and coursework was introduced. Over a two-year period, students produced a few pieces of extended work, each taking about two weeks of curriculum mathematics time.

(It is easier, when introducing a new assessment, to design something closer to your declared objectives than when modifying a well-established examination. The degradation of the CSE began when the admirable wish to have a score-equivalence with the more-prestigious O-level (CSE Level 1 = O-level Grade C), led to demands that the two examinations look much the same. The functional usefulness and real world focus of CSE Mathematics gradually disappeared into the familiar academic forms of O-level)

CSE was scored by the students' own teachers, using a well-defined set of criteria. Standards were monitored through a process of "consensus moderation", where groups of teachers came together to review samples of each other's student work – first within each school, then across a number of schools, chaired by an external moderator. Some remarkable work was produced (though more in other subjects than in Mathematics); much was mundane.

As a result of the required coursework, all students became experienced in producing substantial pieces of coherent work, and refining its accuracy, thoroughness and communication – qualities well-beyond those assessed in timed written examinations. (There were these as well) The change in the implemented curriculum was clear, as was the professional development provided by the consensus moderation process, where teachers shared their experience and showed each other the standard of student work that

[1] Certificate of Secondary Education

[2] General Certificate of Education, Ordinary Level – as opposed to the Grade 12 Advanced Level

could be achieved. Equally, there is no clear evidence of much influence on how the rest of the curriculum, preparing students for the timed exams, was taught and learned.

Coursework has remained a component in much mathematics assessment in Britain. Its pervasiveness is reflected in the complaints of teachers at the curriculum time that is absorbed by it. Its quality has tended to degrade over the years – the projects becoming more routine and similar, from student to student and from year to year. (We shall discuss the pressures that lead to this near-universal degradation process, and how it may be avoided, below)

We have described this example in some detail, as typical of the changes in classroom practice that follow from introducing an assessment component that is, at least initially, well-engineered. Broadly, the required change happens, but with limited influence on the rest of the implemented curriculum. As always, there may be some unintended consequences, which good design and careful systematic development can minimize.

For timed examinations there is some limited evidence on the effects of changes on classroom practice, apart from the research studies reported next. For example, under the *Testing Strategic Skills* project, when the Joint Matriculation Board introduced a new task type, year-by-year, into its O-level examination, in Mathematics. the sales of the associated modules of teaching and professional development materials (Shell Centre 1984, 1986) represented a substantial proportion of the schools that took the examination – far exceeding sales of similar materials which lacked a specific link to high-stakes assessment. However, what the teachers did with the material was not observed, though informal feedback suggests that they were used – performance on the new task types also improved.

Two research studies

The two studies that address the WYTIWYG question most directly are both from Australia but the importance of high-stakes assessment there is similar to the US. The first is on Mathematics in the upper stages of high school, the other first year undergraduates in Sociology. Other studies (see eg Stecher and Hamilton, 2004) look at similar issues from a somewhat different viewpoint, where the test remains the same but the accountability pressures are changed, and come to similar conclusions on WYTIWYG; however, this approach overlooks the opportunities for improvement.

A comparison of two states' assessment procedures and learning practices

Barnes, Clarke & Stephens, 2000

To discover whether a systematic probe of the facts supports this received wisdom, a study was conducted in two Australian states, Victoria and New South Wales, which have similar populations, but differ in their curriculum specifications and their assessment practices.

For the Victoria Certificate of Education (VCE) at the time of the study, the curriculum documents specified three *work requirements*, each of which was to occupy at least 20% of the class time, and on each of which students had to perform adequately to pass. These were:

- *Problem-solving and modeling* – the creative application of mathematical knowledge and skills to solve problems in unfamiliar situations, including real-life situations;
- *Skills practice and standard applications;*
- *Projects* – extended, independent investigations involving the use of mathematics.

In Year 11 (~Grade 11) schools were responsible for carrying out assessment tasks based on these work requirements. In Year 12, there were three *Common Assessment Tasks.*

One of these was a centrally-set, school-assessed task taking two or more weeks, in and out of class; this took the form of an investigative project or a problem-solving task, involving a written report.

In New South Wales (NSW), curriculum documents included a *Statement of Principles* which discussed the nature of mathematics learning, investigation, learning from others, and communication. Syllabuses for the early secondary years emphasized problem solving, investigation and communication, but there was no *requirement* at any level to incorporate such material in school assessments. There were external examinations in Years 10 and 12, based on traditional types of question, including multiple-choice. Guidelines indicated that schools might, if they wished, use forms of assessment other than timed written tests. To ensure comparability, in-school assessment results were standardized by reference to the written examination – so time spent on such things could not increase scores, but reduced time preparing for the written examinations.

To ascertain the impact of these differences in assessment on classroom practice, the study used documents, teacher questionnaires and interviews (but no direct study of student work). The documents included official curricula and guidance, school departmental documents, teachers' worksheets and other classroom material. The questionnaires and interviews focused on the teachers' perceptions of how the external assessments influenced their practice. In all 11 Victorian schools, the course documents and class materials showed substantial incorporation of problem-solving and investigative tasks, and most teachers reported spending between 10 and 30% of time on these aspects.

In their ratings of which aspects of teaching were 'highly important', there were big differences between Victorian and New South Wales teachers. These were in the valuation of:

- presenting problems requiring a range of problem-solving techniques;
- students developing investigative skills;
- using problems to develop mathematical skills;
- developing report-writing skills;
- undertaking extended mathematical activity and open-ended activities;
- regular completion of student mathematical journals.

There was also evidence of the use by teachers in lower grades, not directly involved in teaching for the for the VCE, of project reports and problem-solving reports for assessment, indicating a spill-over effect of the influence of the examination style that reached right down the years of secondary schooling.

The parallel study of NSW teachers showed that their curriculum documents, and their reports to parents, focused strongly on examination results. Their comments on problem solving and investigation emphasized difficulties and costs; 76% mentioned that covering the syllabus left little or no time for such activities. School assessments were based on examination results. The teachers felt that the main external examination, the Higher School Certificate (HSC), tended to make them more content-oriented in their teaching even in the junior years. The Year 10 test also caused pressure to complete the syllabus in time for the test. This test did not carry any high stakes for the students, but teachers nevertheless stressed its importance. The use of student mathematical journals, extended and open-ended mathematical activities and the development of report-writing skills received very little endorsement from these NSW teachers. Indeed, it appeared that the NSW teachers had little knowledge or awareness of such activities.

The striking outcome of this study is that two states whose high-stakes assessments embody such contrasting values showed corresponding contrasts in the instructional practices, in spite of strong similarities in their curriculum specifications and guidelines. In both states, teachers make extensive use of sample questions, and examination papers

from previous years, to interpret the syllabus, to guide their teaching, and to prepare students for their assessments.

We have thus, in this study, empirical evidence of the powerful influence of assessments on teaching, with some details of how the mechanisms operate. WYTIWYG is not just 'received wisdom'

A study of student learning under two assessment regimes

Scouller, 1996

This study of first year students in a sociology course at the University of Sydney shows that assessment is particularly significant in influencing students' approaches to their studies, and the content of their learning. First-year students were asked to reflect upon the way they prepared for and perceived two assessments within the same course – an assignment essay and their short answer examination. A three-part questionnaire elicited:

- students' responses on their learning strategies and motives (classified as either deep or surface);
- their perceptions of the levels of intellectual abilities being assessed by these two tasks (classified either as lower or higher);
- their preference for either one as the method for assessing their learning of the course.

Results indicated distinctive patterns of preparation and perceptions according to assessment method. Students were more likely to employ surface strategies and motives when preparing for their short answer examinations and deep strategies and motives when preparing their assignment essays. They were also more likely to perceive the short answer examination as assessing lower levels of intellectual abilities and the assignment essay as assessing higher levels of cognitive processing.

The degradation of examinations

While the evidence outlined above shows the potential power of high-stakes assessment as an engine for improvement in the implemented curriculum (and why such improvement is so difficult to achieve when the assessment is more narrowly focused), a cautionary note is important.

Many improvements that have been introduced into high-stakes assessment in the past have, over a period of years, lost many of their qualities. Mathematics tasks that used to require longer chains of reasoning get broken down into smaller sections. Coursework that initially produced interesting, varied and even original extended pieces of autonomous student work gradually becomes more and more routine and imitative. Indeed, recent modifications to the assessment requirements in Victoria suggest some degree of degradation – a mix of tests and smaller investigations, now set by the teacher with guidance from the curriculum authority, replacing the former quite substantial centrally-set school-assessed tasks.

Why does degradation happen? Though the direct evidence is sparse, there are always pressures that tend steadily to degrade an established assessment system. They include:

- teacher expectation that papers will contain only minor variations on familiar task types that have been practiced in class, particularly when an examination has been running for some time;
- limited opportunity for question setters to develop new and worthwhile tasks beyond the routine, partly because of the lack of opportunity for trialling and improving draft tasks;

- teacher insecurity produced by anything non-routine, though tackling non-routine problems is at the core of doing mathematics – and it *can* be fairly assessed;
- a concern for 'fairness' which makes reliability the key goal and validity across the range of performance goals a secondary consideration;
- commercial pressures on assessment providers as systems choose the least expensive tests – and, under pressure for "good results", search for the easiest examinations in each subject.

This list could go on.

This downward pressure on assessment quality (as measured through match to system standards and their performance goals) can be resisted but only by ensuring that there is an active *engine for improvement* within the assessment design process, re-injecting those elements that tend to get squeezed out – new tasks, checks on alignment with performance goals, and so on.

Strengths

This evidence provides the core of a powerful case that system-wide improvement will only happen if the assessment matches the curriculum and professional development goals – something that is likely to be new to many school boards, and some superintendents.

Likely challenges

More authentic assessments may be resisted on account of cost.

Teachers may be put off by unfamiliarity with such assessments.

Parents may also find them unfamiliar though, after explanation and some reflection, they often come to welcome them.

References

Barnes, M, Clarke, D, and Stephens, M, 2000. Assessment: the engine of systemic curricular reform?, *Journal of Curriculum Studies,* 32 (5), 623-650.

Scouller, K, 1996 Influence of assessment method on students' learning approaches, perceptions and preferences: The assignment essay versus the short answer examination, in *Different Approaches: Theory and Practice in Higher Education,* Proceedings HERDSA Conference 1996. Perth, Western Australia, 8-12 July. http://www.herdsa.org.au/confs/1996/scouller.h

Shell Centre, 1984 *Problems with Patterns and Numbers*, Manchester: Joint Matriculation Board, reissued Nottingham: Shell Centre Publications 2000 *www.mathshell.com*

Shell Centre, 1986 *The Language of Functions and Graphs*, Manchester: Joint Matriculation Board, reissued Nottingham: Shell Centre Publications 2000 *www.mathshell.com*

Stecher, B and Laura Hamilton, L, *Putting Theory to the Test*, http://www.rand.org/publications/randreview/issues/rr.04.02/theory.html