EDUCATIONAL DESIGNER
JOURNAL OF THE INTERNATIONAL SOCIETY FOR DESIGN AND DEVELOPMENT IN EDUCATION

High-stakes Examinations to Support Policy

Design, development and implementation

Paul Black, Hugh Burkhardt, Phil Daro, Ian Jones, Glenda Lappan, Daniel Pead, and Max Stephens

Abstract

How can we help policy makers choose better exams? This question was the focus of the Assessment Working Group at the 2010 ISDDE Conference in Oxford. The group brought together high-level international expertise in assessment. It tackled issues that are central to policy makers looking for tests that, at reasonable cost, deliver valid, reliable assessments of students' performance in mathematics and science with results that inform students, teachers, and school systems.

This paper describes the analysis and recommendations from the group's discussions, with references that provide further detail. It has contributed to discussions, in the US and elsewhere, on "how to do better". We hope it will continue to be useful both to policy makers and to assessment designers.

Executive Summary

What makes an exam "better"? High-stakes testing has enormous influence on teaching and learning in classrooms – for better or for worse. Teachers give high priority to classroom activities that focus on the types of task in the test. This is understandable, indeed inevitable – after all, their careers are directly affected by the scores of their students on these tests, the official measure of their professional success. Yet this effect of testing on teaching and learning seems to be ignored in the choice of tests by policy makers, who see tests only as measurement instruments. Driven by pressures for low cost, simplicity of grading and task predictability, current tests have a narrow range of item types that does not reflect the breadth of world-class learning goals as set out, for example [1], in the Common Core State Standards for Mathematics (CCSS) or, indeed, in many of the

state standards that CCSS is replacing. Yet, conversely, high quality exams can help systems to educate their students better. Good tests combine valid and reliable information for accountability purposes with a beneficial influence on teaching and learning in classrooms – i.e. they are tests worth teaching to.

How do we get better assessment? This paper, building on substantial experience worldwide, sets out the essential elements of an assessment system that meets this goal, taking into account the necessary constraints of cost. In brief, this entails:

- Planning assessment, including high-stakes testing, as an integral part of a coherent system covering learning, teaching and professional development, all focused on the classroom.
- Treating assessment as a design and development challenge, first to introduce high-quality instruments which serve both formative and summative purposes, then, later, to counteract the inevitable pressures for degrading that quality.

The task of creating an assessment system to help guide and enhance teacher practices and students' learning should include the following steps:

- Create a pool of tasks from a variety of sources each exemplifying
    - Research-based understanding of learning and performance
    - Creative designs
    - Refinement through trialing in classrooms
    - Feedback from teachers and others;
- Establish authoritative committees, independent of the test vendors, with the needed expertise to select tasks from the pool so as to balance the tests across performance goals, as summarized in the standards;
- Involve teachers in both the test design and the scoring processes, using the potential of test design and of scoring training as powerful modes of professional development, with built-in monitoring to ensure quality and comparability across and between teachers and schools; and
- Grow human capacity by providing training for designers of curriculum and of assessment.

These recommendations are amplified and justified in the following sections.

1. The roles of assessment in education systems

Good educational systems must have the capacity to evolve over time. Testing systems must also have this capacity, both in relation to their purposes and the actual assessment instruments that are created. Given the more rigorous demands on teaching and learning that have become accepted internationally, exemplified by the US CCSS, test validation requires a concomitant rigor with a broad range of strong evidence.

To achieve the corresponding educative value, high quality exams will require radical change in system design. The extent of the challenge may be seen by comparing familiar current tests with this extract from CCSS, reflecting international standards for mathematics:

> Mathematically proficient students understand and use stated assumptions, definitions, and previously established results in constructing arguments. They make conjectures and build a logical progression of statements to explore the truth of their conjectures. They are able to analyze situations by breaking them into cases, and can recognize and use counterexamples. They justify their conclusions, communicate them to others, and respond to the arguments of others. They reason inductively about data, making plausible arguments that take into account the context from which the data arose. Mathematically proficient students are also able to compare the effectiveness of two plausible arguments, distinguish correct logic or reasoning from that which is flawed, and—if there is a flaw in an argument—explain what it is. Elementary students can construct arguments using concrete referents such as objects, drawings, diagrams, and actions. Such arguments can make sense and be correct, even though they are not generalized or made formal until later grades. Later, students learn to determine domains to which an argument applies. Students at all grades can listen or read the arguments of others, decide whether they make sense, and ask useful questions to clarify or improve the arguments.

Examination design should reflect as far as possible the full range of goals of the mathematics curriculum. Anything less would not be acceptable as valid implementation of the intentions of the standards.

The examination systems which will be developed should incorporate an auditing mechanism for checking how well the assessment practice is realizing the intentions. Such a mechanism should identify problems – for example, that the current system of curriculum and assessment is not preparing students for the higher levels of mathematical thinking and reasoning embodied in CCSS, or other international standards.

Policy documents for school mathematics often point to the importance of mathematical proofs, mathematical modeling and investigations in achieving a balanced curriculum. However, these remain "paper expectations" and will receive little attention in the classroom unless there is a suite of required student assessments that will make clear to teachers the need for instruction to pay attention to these aspects of performance.

Most current assessment systems fail to give students opportunity to show the range of desirable performances on educative exams in an environment that supports them in raising their own level of performance. The system sends score reports back to district and school offices but the information most often does not match the purposes of the various users.

The district, which wants a limited amount of data that is valid and reliable, believes that these reports give a fair picture of the level of achievement of students, teachers and schools. Yet the multiple choice tests on which they are based, consisting of short items, assess only fragments of mathematics, not students' ability to use their mathematics effectively, to reason mathematically about problems [2].

Teachers may learn a bit about the overall level of students from these reports, but not the kind of information that would enable them to help their students to raise their level of performance. Students may receive a score and a placement in the class as a whole, but with little to no information that could help them understand what higher quality performance entails.

What school systems need is a valid and reliable overall picture of performance of students, classes and schools. In this regard, validity requires tests that are balanced across the performance goals, not just testing those aspects that are easy to test.

What teachers and students need is detailed differential information on their strengths and weaknesses, accompanied by instructional ideas to help build on strengths and to remediate weaknesses. In this regard, valid information can only come from assessment of performance on mathematical tasks that demand substantial chains of reasoning. Most of this feedback needs to be formative, detailed, and timely enough to inform action.

Knowing common patterns of mistakes, and locating student performance along the underlying developmental continua of learning in a given class can help teachers plan remediation, as well as change their classroom teaching practices for future students. To stimulate teachers toward this goal, rubrics for scoring and sample papers for typical responses at each level of performance on the exam can promote changes in classroom practices. Reports that include typical responses of students at different scoring levels can be used in discussions with students to

provide further learning opportunities. Separate scores on different dimensions of tests, along with individual item results, can help teachers improve their practice. Students should see where their performance lies along a progression of learning, so that they also understand the kinds of responses that would be classified as high quality work.

The goal is to make the examination system educative to students, teachers and parents. Timeliness is central – if the exam results, scoring rubrics, and sample papers are not returned promptly, the window of interest, engagement, and learning for teachers, parents, and students will have closed. Teachers and students will have moved on to other parts of the curriculum and have no enthusiasm for feedback that is not currently relevant. Insofar as teachers' can be trained and then trusted to do the scoring themselves or with close colleagues, this speedy response can be secured more easily.

It is clear from the above that high quality assessment is an integral part of a coherent education system, linked directly to the improvement of teaching, learning and professional practice. This should not be a surprise; all complex adaptive systems are like this, with feedback designed to enhance every aspect of system performance. This is the strategic design challenge. In the following sections we describe how it can be, and has been, met.

This strategic view puts the issue of cost in perspective. The cost of various kinds of assessment must be seen in terms of the overall cost of educating a student, approaching 10,000 US dollars per year. Is assessment cost effective in delivering improved education? To make it so is primarily a design and development challenge.

2. Design principles and practices for summative examinations

In this section we outline the principles that guide the design of examinations, intended primarily for summative purposes, that aim for high quality, and the practices that enable these principles to be realized. We start with the principles – the criteria that set out what we mean by high quality. They are fairly non-controversial, probably shared by most knowledgeable people – at the rhetorical level, at least. They are often neglected in practice.

Validity

The key here is to "assess the things that you are really interested in" – to assess students' progress towards the educational aims, expressed in performance terms, for whatever purposes the outcomes of the assessment are to be used for.

While this may seem so obvious as to be hardly worth stating, this fundamental principle is widely ignored in the design of examinations where, for example, the tasks are confined to elements of performance that are easy and cheap to assess.

Another common failure is to specify the outcomes of interest in terms of a simple model of performance in the subject – for example, a list of some elements of performance (usually individual concepts and skills) – and testing these fragments separately. Why is this problematic? Because there is no assessment of students' ability to integrate these elements into the holistic performances that are really of interest – for example, solving substantial problems.

In seeking validity, certain questions deserve particular attention in test design:

- *Inferences:* Can the users[3] make valid inferences about the student's capabilities in the subject from the test results achieved by that student?
- *Evaluation and Decision:* Can users evaluate the results and use them in making decisions, including pedagogical ones, with confidence that the results are a dependable basis, free of bias effects and reflecting a comprehensive interpretation of the subject's aims?
- *Range and variety:* Does the variety of tasks in the test match the range of educational and performance aims – as set out in, for example, CCSS or other specification of the aims of the intended curriculum?
- *Extrapolation:* Does the breadth and balance of the domain actually assessed justify inferences about the full domain, if the former is a subset of the latter?
- *The effects the test has on what happens in classrooms:* Both common sense and data from observations show that, where there are high-stakes tests, the task types in the test dominate the pattern of learning activities in most classrooms. Does this influence represent the educational aims in a balanced way?
- *Is this a "test worth teaching to"?* Given this inevitable effect on classrooms, this question summarizes a very important criterion of validity.

Many high-stakes tests do not score well on these criteria. Validity for the users is sometimes justified by using evidence of correlation with other measures: this simply calls in question the validity of those other measures in relation to the aims of assessments being justified. Very often, the effects on classrooms are simply ignored – tests are seen as "just measurement". Validity of assessments may also be justified by redefining the educational goals to be what the test assesses. The harmful effects are exacerbated if the curriculum aims are expressed in very vague terms: then the test constructors become the arbiters who translate vague aspirations into specific tasks that convey an impoverished message to teachers.

Reliability

Reliability is generally defined as the extent to which, if the candidate were to take a parallel form of test, on a different occasion but within a short time, the same

result would be achieved. Clearly, reliability is a necessary condition for validity, but not a sufficient one – some argue that it should be treated as a component of validity. For a more detailed account, which give a full discussion of criteria, all of which are discussed here see Stobart (2001). Some of the main threats to reliability are:

- *Occasion variability:* The student may perform at different levels on different days.
- *Variability in presentation:* The way a task is explained may sometimes be inadequate, and the conditions in which the assessment is attempted may be inappropriately varied.
- *Variations in scoring:* There may be poor inter-rater or intra-rater consistency – while simple responses to short items can be scored automatically, trained people are needed to score responses to complex tasks. Weak scoring can also threaten validity if the scoring protocol is too analytic, or too holistic, or fails to capture important qualities of task performance. Where raw scores are used to determine grades, variations in the setting of grade boundaries may also be problematic.
- *Inappropriate aggregation:* Variations in the weights given to different component tasks will threaten both reliability and validity, as will inconsistencies between the scoring criteria for the different tasks included in the aggregation.
- *Inadequate sampling:* A short test will have lower reliability than a longer one, other things being equal, because a smaller sample of each student's performance will have greater fluctuations due to irrelevant variations between tasks. To narrow the variety of tasks may produce an unacceptable reduction in validity, so more extensive and longer assessments may be needed to cover a wider variety of performance types with the same reliability. A recent US study finds that a broad spectrum test of performance needs to be four times as long as a short item multiple choice test for the same reliability – close to the few hours of examinations common in other countries (Black & Wiliam, 2012). However, if the aggregated tasks are too diverse the result may be hard to interpret, i.e. weak in validity.
- *Variation between parallel forms:* For any assessment of performance on non-routine tasks, variation from form to form among such tasks is essential to ensure that they remain non-routine; this can offset 'teaching to the test' and stereotyping through repetition of the same tasks from year to year, but it may also introduce irrelevant variability.

- *Variation in the setting of grade boundaries:* Where raw scores are converted into grades, the score at which each grade boundary is fixed can have a very marked effect on those with scores close to the boundary. The influence of this feature depends on the number of grades in relation to the shape of the score distribution.

It will be evident from the above that there is strong interaction between reliability criteria and validity criteria. The key principle is that irrelevant variability should be minimized as long as the methods used do not undermine validity – there is no value in an accurate measure of something other than what you are seeking to assess. Mathematics assessors tend to be proud that their inter-scorer variation is much lower than, say, that in the scoring of essays in History or English; however, this is largely because current math tests do not assess holistic performances on substantial tasks.

Poor understanding of test reliability causes problems. Ignoring the fine print in all test descriptions, users are inclined to view test results as perfectly accurate and take decisions on that basis. The likely degree of variability in test scores should be published, but the forms and language of communication have to be chosen with care: in common parlance error conveys a judgment that somebody made a mistake, and to say that a result is unreliable may convey to many a judgment that it is not fit for the purpose (He, Qingping & Opposs, Dennis, 2010). Thus, it is important to distinguish between 'error', such as mistakes in scoring, and other sources of 'variability' which are either in principle unavoidable or can only be avoided by unacceptable means e.g. if twelve hours of formal testing were required to improve the sampling range.

Capacity for evolution

No test is perfect; some are not even adequate for their declared purpose. Further, the educational aims will change. For high quality assessment, it is essential that tests can grow along with system improvement.

This is particularly important when there are new initiatives, like the current one in the US based around the Common Core State Standards. (It is unlikely that current school systems can immediately absorb tests that fully reflect the aims of these standards; it would be unfortunate if limited first versions were regarded as a long term solution.)

Equally an assessment system should be designed to counter degeneration under inevitable system pressures for:

- *Task predictability:* High-stakes tests make teachers, understandably, feel insecure. They seek tests that are highly predictable, built of tasks that will be routine exercises for their students. However, such tests do not assess the ability to tackle non-routine problems – an important educational goal.
- *Removal of weaknesses:* Tests of higher-level elements of performance, which present greater design and development challenges than routine tests, will have engineering weaknesses, particularly in the early years. (The weaknesses of familiar tests are ignored.) There will be pressures to remove the novel elements, though the weaknesses could be addressed in other ways [4].
- *Cost containment:* Although the costs of high-stakes assessment are a very small proportion of the overall costs per student, they are often regarded as a separate accountability budget line, ignoring the potential contribution of more valid examinations to enhancing the quality of education.

To support evolution and succeed in resisting these downward pressures requires an active *engine for improvement* within the assessment system [5].

Turning principles into practices

What is needed to turn these principles into a system that delivers high quality examinations? Two things stand out:

- *Variety of evidence:* Validity demands a broad range of types of evidence, based on a wide variety of task types; most current systems draw on too narrow a range.
- *System design:* Improvement requires radical change to assessment systems; current systems are not adequate for the design and development of high-quality tests

How do we do better? Both problems need to be tackled together. There are well-established models that have been used around the world. The key features in achieving high-quality seem to involve three elements, each independent but interacting with each other:

Task design, creating a pool of tasks from a variety of sources, each using:

- Research-based understanding of the learning and performance goals;
- Creative designers, with substantial experience in the design of tasks that, taken together, cover the spectrum required;
- Refinement through trialing with students in classrooms, with analysis of student responses and comments from students, teachers and others.

This process, and the design skills involved, is much closer to the design and development of classroom materials than to current test development. For many current systems, achieving this will involve, as part of system improvement, a substantial program of designer training.

Task selection and test balancing, under the aegis of an authoritative committee. It is in the balancing of the overall assessment that those responsible to the community identify and weight the elements they want to value. This is a quite different role from creating the range of opportunities that is the responsibility of broad-spectrum task design; they should be kept independent.

Management of the testing process, by a competent agency, should again be independent of the task design and test balancing process. We shall say little about this, because there are testing agencies around the world which can handle this aspect well.

As well as formal examinations, other components are important for high-quality assessment. The need to produce the broad range of types of evidence that may be required cannot be met solely by tasks set in the controlled conditions of formal testing. Notable examples are tasks requiring extensive investigation by students, and tasks involving group collaboration. For many such tasks, the scoring has to be done by the school's own teachers. To achieve high quality in these assessments, it is necessary to ensure validity in the tasks presented to students, some uniformity in the conditions in which they are attempted, and validity in the criteria by which teachers' make their assessments. In addition there have to be procedures in place to ensure comparability in these key features by both inter-school and intra-school assessment and monitoring development programmes (Black, Harrison, Hodgen, Marshall & Serret, 2011). Such developments require sustained effort but several countries or states have set up such systems and tested them to ensure that the assessments produced are comparable in quality to those of external tests, and valuable as complementing in specific ways the results of such tests (Black 2010, Hayward, Dow & Boyd 2008).

To summarize, we need:

- Much better examinations
- Regular curriculum embedded tasks, each representing clear learning outcomes
- Some assessed group work with fair credit assignment.

All this needs to be managed by a system that will ensure the credibility of the results, and that all components are valued. (For example, a rule like "50% of the grade must come from grading 'revised' work", where 'revised' means submitted for feedback but not graded, revised after feedback and then graded.")

There are many other sources of evidence about the positive effects of ways to involve classroom teachers in the system. It brings teachers' expertise into assessment and, complementing this, gives them effective professional development in various ways. It builds their understanding of the system, and of the educational aims of the examination. It builds their confidence in the system, while also helping them to form a positive relationship between summative and formative assessment processes and the use of both in harmony to enhance learning. At the same time, there is ample evidence to show that, with care, a system can be designed to ensure that teachers' involvement does not reduce the reliability of the system.

Auditing

For such a system to move forward consistently it needs mechanisms for checking how well the assessment practice is realizing the intentions. These include:

- Identifying matches and mismatches between the intentions and their realization in the assessment
- Fixing the mismatches
- Preparing to improve the assessment by the development of collections of new types of task
- Working actively for improvement at system level

We expand on this in Section 5.

Inspiring examples

Nuffield A-level Physics in England included work produced under formal test conditions, and other work produced in more flexible situations (Pellegrino *et. al.*, 2001) . In the former category were:

1. A multiple choice test of 40 questions (75 minutes) 20%
2. A test of 7 or 8 short answer questions (90 minutes) 20%
3. Questions on a passage from outside the syllabus to test ability to deploy physics knowledge to new things (150 minutes) 24%
4. A practical problems test in a laboratory: candidates went around eight "stations" to make measurements, suggest possible investigations and so on (90 mins.) 16%

In the second category were:

5. A project essay involving researching and writing about a topic chosen by the student - done over about 2 weeks in normal school time 10%
6. An open ended investigation on a different topic for each student assessed by the teacher who sent in samples, done over about 2 weeks in normal school time 10%.

This pattern of diverse assessment tools worked well in the normal context of an A-Level end-of-school examination in science.

The VCE Mathematics Examination in Australia in the 1990s involved two 90 minute end of year exams. Paper 1 was multiple choice and Paper 2 was extended questions of about 20 to 30 minutes each. During the year there was an investigatory project developed by a panel of experts and of high quality. There were however concerns about the amount of help students received due to the high-stakes nature of the assessment. This was addressed using a post-investigation test which was cross-referenced with the log-book kept by a student during the investigation. In the current system external test and internal assessments by schools each contribute 50% to the total scores, with careful studies of inter-correlations between the two scores to reveal and explore any anomalies (Stanley *et. al.*, 2009).

There is a fuller description of these and other successful examples in section 4 of Burkhardt (2009).

The 'Novice-apprentice-expert' model

This model, being developed in the US in response to CCSS, is one of a number of approaches that may help acceptability. They involve different Mathematical Practices (MPs):

- *Novice tasks* are short items, each focused on a specific concept or skill, as set out in the standards. They involve only two of the mathematical practices (MP2 – reason abstractly and quantitatively; MP6 – attend to precision), and do so only at the comparatively low level that short items allow.
- *Apprentice tasks* are substantial, often involving several aspect of mathematics, but structured so as to ensure that all students have access to the problem. Students are guided through a "ramp" of increasing challenge to enable them to show the levels of performance they have achieved. While any of the mathematical practices may be required, these tasks especially feature MP2, MP6 and two others (MP3 – construct viable arguments and critique the reasoning of others; MP7 – look for and make use of structure). Because the structure guides the students, the mathematical practices involved are at a comparatively modest level.
- *Expert tasks* are rich tasks, each presented in a form in which it might naturally arise in applications or in pure mathematics. They require the effective use of problem solving strategies, as well as concepts and skills. Performance on these tasks indicates how well a person will be able to do and to use mathematics beyond the mathematics classroom. They demand the full range of mathematical practices, as described in the standards, including: MP1 – make sense of problems and persist in solving them; MP4 – model with mathematics; MP5 – use appropriate tools strategically; MP8 – look for and express regularity in repeated reasoning.

The three types aim to present comparable difficulty, but with a different balance of challenge – largely technical for novice tasks, more strategic for the others. "Easy expert" tasks, allowing a variety of approaches, are at the heart of mathematical literacy [6].

System issues in making it happen?

Designers need to understand the constraints and affordances of the current system, so as to identify which constraints can be pushed and which are immovable.

Policy makers are understandably reluctant to accept that their examinations are inadequate – any change in high-stakes examinations provokes anxiety and a correspondingly strong public reaction. So though the assessment of subject practices, rather than just content knowledge, meets resistance, other subjects can show the way. For example, English (Language Arts) has long assessed writing performances as complex as those proposed for mathematics.

Policy makers need to be convinced they will get a (societal) return for the price of improved high-stakes assessment – both financial and political. Rational

arguments can be made but creating broad acceptability is a key ingredient of success. For example, Nuffield A-level physics required convincing key stakeholders (the physics community and university admissions people) and, as a voluntary alternative for schools, did not directly threaten the status quo.

The longevity of innovations that do get off the ground is another challenge. In many successful cases, when the system changes inspiring examples can disappear. This emphasizes the need for active *engines for improvement,* discussed above.

3. Scoring, monitoring and reporting

A well-designed assessment gives every student the opportunity to perform – to show what they know, understand and can do in the subject area.

Scoring [7] is the process of evaluating each student's responses – "evaluating" implies a set of values, usually embodied in a scoring rubric that guides the scorer. Society demands that the scoring process is fair to each student's performance, so that two different scorers will give much the same score to the same response; the degree of acceptable variation between scorers varies from subject to subject and country to country. Scoring of essays has much wider variation than scoring short factual answers. The US public likes machine scoring, partly because the score-rescore variation is negligible; there has been less concern that the limits of machine scoring narrow the range of performances that can be assessed, undermining the validity of the assessment. Use of machine-scored essays is increasing, but it too may be shown to limit the range of performances that can be assessed. In the scoring of extended answers, numerical scores are allocated to each part successfully completed. This kind of scoring is often carried out by teams of trained scorers, as is the scoring of rich tasks for which holistic scoring rubrics are used to construct an overall score based on how well the response meets specific performance criteria. As we stressed in Section 2, improving reliability is a valuable goal only if it is not at the expense of validity.

As pointed out in Section 2, variation between scorers is not the only limit on reliability. The variation between scores that the same student gets on supposedly equivalent tests is equally important, and often larger. This test-retest variation is usually mentioned, if at all, only in the technical fine print of a test description [8]; the public continues to believe that tests are accurate and that life-changing decisions taken on the basis of test scores are reliable and fair.

Monitoring processes are used to increase the reliability of scores. Some are simple checking – for example, that scores from student papers have been correctly entered into computer systems. Others are more profound – for example, second scoring by a senior examiner of a sample of papers, or consensus monitoring, where groups of teachers examine samples of each others' scoring to

ensure consistency of standards. We shall discuss monitoring procedures in context below.

Reporting the results may seem straightforward but issues arise there, too. How much detail should be reported? Detailed scores are statistically less reliable than aggregated totals – yet aggregation throws away valuable information. Fortunately, there is a happy trade off here, allowing us to meet the different needs of different users and uses. At one extreme, school systems want a small amount of data that they can believe is valid and reliable [9]. Aggregated scores for classes may provide evidence for teacher evaluation, aggregating further for school evaluation. On the other hand, teachers and students need detailed feedback on responses to individual tasks to guide future work; they are less worried about variability – that a student might have done somewhat better or worse with a similar but different task on a different day. Parents, somewhere in between, want to know about their children's performance in more general terms. How is s(he) doing on numerical skills, on understanding concepts, and on problem solving? All need to know that the information they get is a valid assessment of performance in the subject whether mathematics, science or language arts. So it bears repeating that tests need to assess performance in mathematics or science in a balanced way – an obvious requirement that is far from much current testing.

In summary, different purposes of assessment have a different balance of needs for scoring, monitoring and reporting. Where results are used for formative and diagnostic purposes detailed feedback is needed. By contrast, where the purpose is summative, as in periodic and final course testing, less and less detail is needed, whilst, as the stakes get higher, the need for reliability increases correspondingly.

On the basis of this background, we now outline a set of principles and processes for scoring, monitoring and reporting on examinations that will provide the various groups of potential users the information they need. The key principles are to:

- *Involve teachers* in the assessment processes in various ways as an integral part of their professional practice. This follows from the need to integrate assessment into the processes of education; it is highly cost-effective. Formative assessment for learning in the classroom is built on a mixture of teacher assessment and student self- and peer-assessment. Teacher scoring of their own or colleagues' student tests, with external or consensus monitoring, provides more detailed feedback as well as reliable scores. Teachers can be effective on-line scorers of complex tasks on tests, as indicated by their participation in scoring tests such as the US "advanced placement" examinations in calculus.
- *Use scoring training as professional development,* integrated into the system's improvement program. It is well-established that professional

development activities built around specific student responses to rich tasks are particularly effective, motivating teachers to focus on key issues of performance, content and pedagogy [10].

- *Communicate the scoring procedures used,* including any criteria used, in each completed assessment cycle to schools and teachers in a clear and timely way to help inform future teaching and to advise students.
- *Use a rigorous procedure to monitor the reliability of the scoring*. There is a wide range of methods, and international experience with using them. Senior scorers may rescore a sample of the scoring of each scorer and, if the variation is unacceptable, adjust the scores, or retrain or reject the scorer. An alternative is to insert standard papers in the stream of papers for each scorer and, again, take action where needed. Or one can arrange for groups of scorers to meet and reconcile their scoring, using samples under the guidance of an expert chair.
- *Use the strengths of IT to support the assessment process.* (see the next section)
- *Involve students* in the assessment processes by returning their work to them as quickly as possible, showing them the scoring rubrics and scored samples of student work at various levels on each task. Student self- and peer-assessment, an essential part of formative assessment, is informed by this review of summative tests. (This is one manifestation of moving students into roles normally used by teachers, a design strategy that generally raises levels of learning.)
- *Most important,* apparent difficulties in scoring or monitoring some types of task are rarely a good reason for excluding them, if they are important for the validity and balance of the examination, including its quality as a test to teach to. There are usually adequate ways of handling the challenges that such tasks pose.

There are many examples of effective scoring of high-stakes assessment built on these principles.

4. The roles of IT

There is an understandable enthusiasm in various places for computer-based assessment, although students who have expereienced computer-based tests of mathematics are not always so enthusiastic, (see Pead 2010, pp187-193). It appears to offer inexpensive testing with instant reporting of results. Here we shall look at the power of computer-based systems for the various phases of assessment: managing the assessment system, presenting tasks to students, providing a natural working medium for the student, capturing the student's responses, scoring those responses, monitoring scoring, and reporting the results. We shall see that IT, at least in its current state, is invaluable for some of these purposes, useful for others,

and very weak for yet others.

- *Managing the assessment system:* Computers can be a powerful aid to the processes involved in large-scale assessment, even with conventional written tests. Scanning student responses saves paper shipping and checking. Presenting responses to scorers on screen and collecting their scores, item by item, within a task allows scorers to work quickly, often at home, and collects data for reporting and analysis. Inserting standard responses that check scorer reliability facilitates monitoring. Most large scale test providers use such systems – and, crucially, they present no obvious problems for a wider range of task types.
- *Presenting tasks to students:* The potential gain of presenting tasks on-screen is that it allows a wider variety of tasks to be delivered in an examination setting. Video can be used to present problem contexts much more vividly. Investigative 'microworlds' in science or mathematics can help in assessing the processes of problem solving and scientific reasoning, enabling students to explore, analyze, infer and check the properties of a system that is new to them.
- *Providing a natural working medium for the student:* This is an aspect that is often overlooked but, if students are to be able really to show what they can do, the mode of working in examinations must be reasonably natural and familiar. In language arts or history, where reading and writing text dominate, word processors provide a natural medium for working, and for constructing written responses. This is a medium that is familiar to most students. However, in mathematics and science, where paper and pencil jottings, sketch graphs and diagrams, tables and mathematical notation are a central part of the way most people think about problems, computers are a clumsy and inhibiting medium. Inputting diagrams, fractions and algebra is slow – a distraction from the problem and an unproductive use of test time. The specialized software that is available takes time to learn, implying changes in curriculum, and standards (see below).
- *Capturing the students' responses:* This is straightforward in the case of multiple-choice or short constructed responses (such as a number or a few words). It is problematic for richer, more open tasks for the reasons explained in the previous point. Currently then, the optimum way of capturing student responses to substantial tasks seems to be through scanning their papers.
- *Scoring those responses:* Automatic scoring of student responses to multiple-choice questions and simple, short answer constructed responses to short items is effective and economical. While progress is being made in scoring more complex responses, major challenges remain to machine scoring of responses to complex tasks in mathematics and

science which, generally, involve sketch diagrams, algebra, etc, set out in no particular sequence or pattern. There is an ongoing danger that the administrative attractions of automatic scoring tempts school systems to sharply limit the variety of task types and the aspects of student performance that can be credited – a prime example of the degradation of assessment through sacrificing validity to statistical reliability and cost.

A different, formative role for automatic assessment is to use computers to search for patterns in students' responses that reveal how they are thinking about a mathematical concept(Stacey *et. al.*, 2009). It is a too-complex task for teachers to go much beyond tallying number of items correct and observing major common errors. However, with the right set of questions, a computer can report diagnostic information to teachers that goes well beyond a measure of how much a student knows. Moreover, this information can be provided to teachers and students immediately, ready for input into the next lesson.

- *Monitoring scoring:* We have noted the role of computers in managing and monitoring on-screen human scoring by injecting standard responses from time to time. Computer scoring of essays has been used to alert a second scorer, a valid and less expensive alternative to double scoring all responses. (We know of no comparable development for complex tasks in science or mathematics.)
- *Reporting the results:* Computers are an essential tool for handling and reporting data for large scale assessments. However, their limitations in the range of data they can capture mean that there is currently no substitute for returning responses to teachers and students, on screen if need be.

  Most commercial computer-based assessment systems offer extensive summary reports and statistical analyses of scores. These are popular with school management, and are a major selling point. The ability to return scored papers to students is less common though not, in principle, impossible.

For designers of a high-quality assessment system, the principle is clear: Use IT for those things where it is strong and avoid it for those where it is weak. Look skeptically at the enthusiastic claims for computer-based testing and scoring systems, especially where their warrants for success come from other subject areas, and ask *whether they can assess the full range of types of performance in mathematics required by CCSS. Test their assertions by asking them to score some real samples of student work on complex tasks.*

Sophisticated testing using batteries of multiple choice questions can capture a large body of evidence from each student. “Adaptive testing” improves this process by selecting the next question based on previous answers. This can be valuable as

part of the assessment regime, particularly for "diagnostic testing" and rapid coverage of the content curriculum, but currently it cannot test a student's ability to autonomously tackle a substantial, worthwhile mathematical problem, requiring an *extended chain of reasoning,* without being led step-by-step through the solution with strong hints as to which mathematical technique to apply at each step.

The danger, though, is that economic pressures will drive computer-based assessment to deliver what is cheap and easy: multiple-choice and short constructed answer tests with the same narrow, fragmented focus that makes so many current 'mathematics' tests invalid as assessments of mathematics.

Wider implications of IT

We conclude this section with some broader questions that need robust answers before the use of IT in the assessment of mathematics is extended. Since they imply changes in both curriculum and standards as well as assessment, they go beyond the main focus of this paper. So we will be brief.

The limitations in the usefulness of computers in assessing mathematics is ironic because computers are central to doing mathematics outside the classroom in everything from simple business calculations to research in many subjects, including pure mathematics. This is not yet reflected in schools where computers and calculators are currently a useful supplement, rather than a powerful replacement for traditional routine procedures. Current curricula and tests mean that most students lack any fluency in the use of spreadsheets [11], computer algebra systems, graphers, dynamic geometry packages [12] and (the ultimate mathematical computing tool) programming. These tools would enable them to realize the power of the computer to develop and support their mathematical thinking. But these aspects of mathematics are not yet integral to most curricula, or to CCSS. This suggests the following questions for the future:

- *Can computer-based assessments incorporate the authentic use of computers as a mathematical tool?* If students are fluent in the use of spreadsheets and the other tools just mentioned, then the computer will become a more natural medium for working, and assessment tasks can be set, to be answered using these authentic mathematical computing tools. This will require changes to the taught curriculum to include the practical use of computers in mathematics – a worthwhile end in itself. Students will learn transferable mathematical IT skills with relevance beyond their school's brand of online test platform.
- *Would this help to improve assessment of mathematics?* Paper-based tasks frequently present a blank space for writing and attract there a range of response elements including sketch graphs and diagrams, tables and mathematical notation. While all of these can be handled on a

computer, students must either be proficient in using these input devices before the test, or the devices must be very, very simple (and hence constrained in what can be entered). There are dangers: presenting the student with the appropriate tool at each stage of the problem (e.g. a graph tool where a graph is expected) can easily reduce an open task to a highly-scaffolded exercise which does not assess the students' ability to autonomously choose and apply the best tools and processes for the problem, or to develop extended chains of reasoning. Inputting answers or other elements of the response is an additional distraction to the students' thinking. There are examples of 'microworlds' that are specifically designed to capture student working but this area of development is still at an early stage.

- *What would you put in an 'essential software toolkit'* that students would be expected to become sufficiently familiar with to use during assessments? We have listed above a range of candidate tools, now used in a minority of schools. (Students will still need desk space for paper and pencil sketching.) For each we should ask: Would these tools embody transferable mathematical computing skills? How, and to what extent should these be introduced into the curriculum in typical schools? This is ultimately a societal decision, as it has been over the many decades it has been dodged.
- *In what ways would standards need to change to encourage the use of such tools in curriculum and assessment?* With the faithful implementation of CCSS still to work on, this question is premature. However, if the gross mismatch between the way mathematics is done inside and outside school is to be addressed, it should be central to the next revision. Meanwhile, we must focus on what can usefully be achieved without such change.
- *How can computers help in formative assessment?* Here we can adopt a more positive tone. Given the recognition in CCSS of the importance of modeling, spreadsheets and other computer tools offer rich possibilities for helping students develop their reasoning skills and mathematical practices. At the simplest level, spreadsheets provide a context for exploring relationships, between variables and with data, that develops insight and provides a 'semi-concrete' bridge between arithmetic and the greater abstraction of traditional algebra as a modeling tool. More generally, the potential for use, in combination, of online discussion boards, tablet PCs and wireless internet access, to foster a collaborative classroom environment opens up new possibilities (Webb, 2010).

It is clear that what emerges from further work on these questions is likely to suggest changes in standards and curricula, as well as in assessment, that belong in the future.

Pead (2010) discusses in further detail many of the points made in this section. An adapted extract from this, describing a project to develop a computer-delivered test using rich problem-solving tasks appears in this issue (Pead, 2012).

5. Steering the system

All complex systems depend for their success on the quality of feedback – in range, depth and timeliness – and the mechanisms for using it to guide improvement in system performance [13]. Here we link the issues discussed above to the roles of assessment in accountability-based management, and how assessment can be designed so as to guide people at all levels towards realizing the system goals.

Steering versus "Are we nearly there yet?"

It has been said that our assessment system seems to have been designed by a 3 or 4 year old in the back seat of a car, repeatedly demanding "Are we nearly there yet?". It should be clear that the under-achievements of US education systems, partly reflected in that of our students in international comparisons, are not going to be solved overnight. Choosing to go for a quick fix guarantees failure. What *can* be done is to establish directions of change and a program of improvement to move education systems forward with deliberate speed in those directions.

Well-designed assessment provides information on direction, not just on distance. In what ways have we improved, and how much? The focus needs to be on the places where student and teacher learning take place – on what happens in classrooms and in teacher professional development. Other system initiatives are effective only insofar as they impact favorably on this "zone of instruction" (Elmore, 1999) . A distribution of spelling scores in her class may provide either threat or reassurance to the teacher; it does not help her know what to do about it. For that, she needs to know which words each student has difficulty with. With mathematics and science, as with spelling, feedback needs to be specific and detailed on all the relevant aspects of performance. Formative assessment of this kind is at the heart of designing assessment to support policy; conversely, it needs appropriate policy to support this kind of assessment.

Well-designed assessment provides directly much of the information needed to steer the system. Are our students learning the full range of concepts, skills and mathematical practices? How effective are they in using mathematics in their problem solving? Formative assessment uses this feedback week-by-week to guide teachers and students along the pathway of progress towards system goals. Summative assessment shows how far each student has progressed in the various dimensions of performance.

In other areas, assessment provides "canary in the mine" indicators of problems in a wide range of instructional practices that need further evaluation. How are

teachers using precious time outside the classroom? Are they working effectively on their professional development? Or are they using the time for routine tasks like grading? How far are our professional development programs being delivered as planned? When they are, what changes are we seeing in teachers' classroom practices? This kind of question can be answered by evaluative studies, provided these are closely linked to the design goals of the programs concerned. Assessment can provide both stimulus and relevant evidence to strengthen such studies.

At the system level, similar questions arise. How well is the program that was successful in the pilot schools spreading system-wide? Do we have realistic support for this, in terms of time and human capital for professional development? Is the pattern of pressure, including the various tests, matched to levels of support that enable most teachers to meet the challenges they face? Is the ongoing funding for improvement at a level that shows that that the leadership is serious and realistic?

At national level, have we provided school systems with the tools and guidance that will enable their leadership and staff to meet the goals?

Assessment and feedback for teaching and learning

Let us summarize the discussion in this paper from the perspective of steering the system.

First, let us ask: What should be taught? What should be learned? In CCSS we have a set of standards which, with associated exemplification, provide explicit guidance – on the practices of doing mathematics as well as the concepts and skills needed.

The next question is: How do you "assess the standards"? Again the answer has changed – constructing test items for each line in the detailed list of standards is not nearly enough. Instead of asking, do the items on a test fit one-to-one to standards, we need to ask: Does the set of tasks in each test have the same focus, balance, coherence and depth of practice as the standards? This shift of viewpoint is important because actually "doing mathematics" involves choosing and deploying multiple elements of the standards for the purpose in hand; focusing on the details alone misses the point.

Learning is a web of progressions, so this is the construct to measure – rather than a "trait" one has more or less of. In designing assessments, we should seek to optimize measurement of growth. For this we need tasks along the progression, not just at the end of the line. This includes capacity to achieve progression within each task not simply between tasks – as students can improve a piece of writing as they progress, so they can improve their analysis and solution of a problem in mathematics.

The influence of high-stakes assessment on what happens in classrooms brings a responsibility on assessment designers to consider motivation. It implies that assessments should measure malleable things that can be taught and learned, rather than fixed traits. It also implies that feedback to students should place less emphasis on ranking and grades and more on guidance, tailored to the needs of each individual, on how that student can improve (Dweck, 2000). Current tests are not based on growth constructs. They are especially haphazard in designing for the construct of growth in the bottom third, but also the middle third of the population. What would the assessment that is designed to measure the growth constructs be like? For example, what do students across the bottom third know, understand, what are their proficiencies? Design an assessment with tasks to characterize the syndromes and detect the growth; right now, the task collections do not do this.

Finally, in reviewing the quality of a test, we must ask the key questions:

> *Is it worth teaching to?* Either way, it will be taught to.
>
> *Is it worth studying for?* Tests either motivate or demoralize.

Multi-dimensional reporting

We have stressed that different users need different kinds and quantities of information. There is no problem with this, provided they are all based on a common view of what is important in performance.

So total scores blur together a lot of different dimensions of performance. It is useful for high-level monitoring of progress by a student, a class or a school but it is inadequate for instruction.

Workers in instruction, teachers and students, need the blur clarified through much more detail – separate scores on different dimensions, task-by-task results and, normally, the return of their own work for review. The reports should emphasize malleability and growth, what has been achieved and what needs to be learned, expressed in concrete terms.

Consequences and audit

No program of innovation plays out as intended – indeed, too often, the unintended consequences have been the major feature of implementation. This should be no surprise. Each of the constituencies of key players (students, teachers, principals, district staff, unions, politicians and parents) have perspectives that will influence how they play their part in implementing the change. It is not possible to predict all this in advance, so success depends on creating a "learning system" that can adjust how it pursues its goals in the light of feedback.

To achieve this, and thus minimize the mismatch of intentions and outcomes, needs regular research on consequences – in particular, how teachers, schools, and districts are, and are not, responding to the assessments. This research will tell us where we are steering. Is it where we want to go? Lack of response or perverse responses may need design modifications. Equally, there may be pleasant surprises that can be built on.

6. Obstacles to progress, and ways of tackling them

Finally, we review the *barriers* to developing a high quality assessment component for a high quality education system. For the processes of seeking improvement, these obstacles are no less important because they are largely a matter of perception. Past experience has shown that they can be overcome, and the benefits of doing so.

Dangerous myths and illusions

Looking at the issues around testing through another lens leads to a realization that current stances toward high stakes assessments encourage some illusions with dangerous consequences. Some have been mentioned above; here we summarize why they are misleading:

- *Tests are seen simply as a measure of student achievement:* Accepting no responsibility for their effect on classrooms has led to narrowing the implemented curriculum so students are only educated in mathematics at the 'novice task' level of the tests.
- *Most attention is given to the statistical properties of the tests* and the fairness of the examination process, with little attention being given to articulating aspects of performance that are actually assessed, as well as their range and balance. Such actions lead to exams that are reliable assessments of fragments of mathematics and to teachers teaching only these fragments.
- *High quality examinations are too costly:* While it is true that they cost more (10-20 US dollars per student test) than machine-scored multiple choice tests (US$1-2), this is still only ~1% of the annual cost of educating a student in mathematics – a small price for invaluable feedback plus professional development.
- *Current tests are inexpensive:* While the cost of the test ($1-2) is small, this omits the great cost of otherwise unproductive "test prep" teaching, which fills many weeks a year in most classrooms (at least $200 per student, lost to learning mathematics)
- *Teacher scoring is unreliable,* and subject to cheating. Provided the training and monitoring are well designed, evidence shows that comparable overall reliability can be achieved. Grading schemes in use across Europe indicate that this is the case.

- *Assessment is a waste of time:* "No child grew taller by being measured." This argument reflects the disdain of many teachers and other educators for testing. (A regrettable by-product is that this community has done little to improve the tests.) Yet the challenge of performance outside the training arena is seen as essential in most fields: in sports, training for the big game; in music, practicing for the concert or the gig; for all of us, going on a course so we can do something better. Further, assessment of the various kinds we have discussed here enhances the teacher's understanding of the strengths and weaknesses of each student – which, evidence shows, is less than many teachers think.

Resources for moving forward

A wide range of useful resources can be found in a number of countries.

Examination systems that embody the various features commended here are, of have been, in general use in various places. There is excellent work on designing rich, challenging, tasks that can be used to examine students' mathematical performance as well as to promote students' mathematical growth and maturing use of mathematical practices

ISDDE has played a role in bringing together designers from many different countries with different views of assessment. These cross-country interactions have already led to design and assessment projects that are taking advantage of varying points of view and experiences to create tasks for learning and for assessment that are of very high quality and that have great potential to "educate" and assess students. One aspect of this work that is especially promising is curriculum-embedded assessments. Such assessment tasks can mirror the external assessment system and give students more detailed formative information about their strengths and weaknesses – far from just a total score. Knowing where you are on a progression toward exemplary work can be a powerful motivator.

While the challenges of rethinking large scale assessments with an eye toward their educative potential for students and teachers seems daunting, the progress that has been, and is being made is encouraging. The fact that the conversation and the work are going on across many countries increases the likelyhood that high-quality educative testing practices will become the norm.

7. Conclusion

The aim being pursued in this study is of fundamental importance if assessment is to be so designed that it encourages the approaches to learning that are needed to prepare students for the future demands of society. Stanley et al. (op.cit.) point out that this calls for a new emphasis on the importance of developing assessment by teachers:

> ...the teacher is increasingly being seen as the primary assessor in the most important aspects of assessment. The broadening of assessment is based on a view that there are aspects of learning that are important but cannot be adequately assessed by formal external tests. These aspects require human judgment to integrate the many elements of performance behaviors that are required in dealing with authentic assessment tasks.
>
> p.31 Stanley et al. (2009)

This is reflected in a broader way in a paper adopted by the European Council of Ministers in June 2010:

> Key competences are a complex construct to assess: they combine knowledge, skill and attitudes and are underpinned by creativity, problem solving, risk assessment and decision-taking. These dimensions are difficult to capture and yet it is crucial that they are all learned equally. Moreover, in order to respond effectively to the challenges of the modern world, people also need to deploy key competences in combination.
>
> (Assessment of key competences: Draft Background Paper for the Belgian Presidency meeting for Directors-General for school education. Brussels: E.U., p. 35 section 6).

Footnotes

[1] We shall use the design challenges that CCSS present as the lead example throughout this paper, but the principles and processes set out here apply to the assessment of mathematics and science quite generally.

[2] It is like assessing basketball players only on their shooting success from the free throw line – relevant but not nearly enough.

[3] Of course, different users may wish to make different interpretations - e.g. some may ask of the successful student “will he/she be able to tackle advanced academic study”, others “will he/she be able to apply what has been learned in a particular work-place environment”. An understanding, clearly communicated, of the aims which an assessment is designed to assess, is essential.