High-stakes Examinations that Support Student Learning: Recommendations for the design, development and implementation of the PARCC assessments

Paul Black, Hugh Burkhardt, Phil Daro, Glenda Lappan, Daniel Pead, and Max Stephens
for the ISDDE Working Group on Examinations and Policy

How can PARCC best achieve its goal of creating better assessments in mathematics? The recommendations in this paper arise from meetings of this Working Group of ISDDE, the *International Society for Design and Development in Education*. The group brought together high-level international expertise in assessment design. 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 – with results that inform students, teachers, and school systems. This paper describes the analysis and recommendations from the group, with references that provide further detail. It is designed to contribute to the conversation on "how to do better".

What is "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. Driven by pressures for low cost, simplicity of grading and task predictability, current tests have a narrow range of types of item that does not reflect the breadth of world-class learning goals, as set out in the Common Core State Standards for Mathematics 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. Such 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*.

In this respect, the Department of Education guidelines for RTT assessment gave reason to be hopeful, as did the consequent proposals from PARCC and SBAC. "Tests worth teaching to" seems to be an idea whose time has come.

How do we get there? 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. It also entails treating assessment as a design and development challenge that entails first introducing high-quality formative and summative assessments, then monitoring the assessments to maintain high quality and to counteract the inevitable pressures for degrading that quality.

Recommendations

1. Create a pool of tasks from a variety of sources, each of which has proven skill in research-based understanding of learning and performance, in creative design, and in the process of refinement through classrooms trials, with iterations of feedback from teachers and others;
2. Establish authoritative committees with the needed expertise who will select tasks from the pool to produce tests that are balanced across performance goals, as summarized in the standards;
3. Select vendors with the capability of administering tests with complex tasks, including human scoring of student responses, with appropriate monitoring to optimize reliability consistent with test validity;
4. Involve teachers in the scoring process, using the potential of scoring training as a powerful mode of professional development, linking high-stakes assessment to curriculum and classroom learning;
5. Grow human capacity by providing training for designers of curriculum and of assessment;
6. Redouble the commitment to ongoing audit, evaluating the content and impact of the new assessments, and taking appropriate actions.

These recommendations are amplified and justified in the following sections. We also present examples in which countries around the world made radical improvements to their assessment systems of the kind being aimed for by PARCC. By providing these examples we hope not only to show that these challenging goals are reachable, but also to provide "leads" which PARCC officials might choose to follow up on to learn more.

Who We Are

Paul Black worked as a physicist for twenty years before moving to a chair in science education. He was Chair of the *Task Group of Assessment and Testing*, which advised the UK Government on the design of the National Curriculum assessment system. He has served on three assessment advisory groups of the USA National Research Council, as Visiting Professor at Stanford University, and as a member of the Assessment Reform Group. He was Chief Examiner for A-Level Physics for the largest UK examining board, and led the design of Nuffield A-Level Physics. With Dylan Wiliam, he did the meta-analysis of research on formative assessment that sparked the current realization of its potential for promoting student learning.

Hugh Burkhardt has directed a wide range of assessment-related Shell Centre projects in both the US and the UK, working with test providers to improve the validity of their examinations. He is a director of MARS, the Mathematics Assessment Resource Service, which brings together products and expertise to help school systems. This often links high-stakes assessment with curriculum and professional development. Currently the team's Mathematics Assessment Project, led by Malcolm Swan, is developing tools for formative assessment and testing to support the implementation of CCSS. Hugh was the founding Chair of ISDDE.

Phil Daro was chair of the writing group that designed the Common Core State Standards for Mathematics. That he was chosen for this role reflects his wide range of experience as consultant and designer at all levels from the classroom to state school systems – for example, as a director of Balanced Assessment for the Mathematics Curriculum, California and American Mathematics Projects, New Standards Project in both Mathematics and ELA, and the current Mathematics Assessment Project. He currently directs the development of a middle school mathematics program inspired by the Japanese curriculum, works on advancing the design and use of leadership tools for change at every level of the educational system, and consults with states and school districts on their accountability systems and mathematics programs. He has served on national boards and committees including: NAEP Validity Committee; ACHIEVE Technical Advisory Group; Technical Advisory Committee to National Goals Panel for World Class Standards; Mathematical Sciences Education Board; and many others. He is Vice-Chair of ISDDE.

Ian Jones is a Royal Society Research Fellow working with the Shell Centre team on the mismatch between intentions and outcomes in the design of high-stakes tests, particularly the UK Grade 10 GCSE examination.

Glenda Lappan, University Distinguished Professor, has led the design of a sequence of middle grades mathematics curriculum projects. She is Director of the Connected Mathematics Project and Co-PI for the NSF-funded Center for the Study of Mathematics Curriculum. She has served as a Program Director at the National Science Foundation and as President of the National Council of Teachers of Mathematics during the development and release of the *NCTM Principles and Standards for School Mathematics*. She is past Chair of the Conference Board of the Mathematical Sciences and Vice Chair of the US National Commission on Mathematics Instruction. In 1996 the Secretary of Education appointed her to serve on the National Education Research Policy and Priorities Board for the US Department of Education. Glenda shared a 2008 ISDDE Prize with Elizabeth Phillips for *Connected Mathematics.*

Daniel Pead has worked in the design and development of educational software, including applets for mathematics education, multimedia products, and computer-based assessment. He directs the IT work of the Shell Centre team, which has included work on a number of assessment projects, notably the *World Class Tests* of *Problem Solving in Mathematics, Science and Technology* for the UK Government. A recurring interest is how to produce computer-based materials which support and encourage good teaching and assessment practice, ensuring that the technology is a means to an end, not an end in itself. He has recently completed a study of computer-based assessment in mathematics. He is the Secretary of ISDDE.

Max Stephens' current research interests are, on the one hand, in student assessment and school improvement and, complementing this, in studying how quite young students begin to move beyond calculation with numbers and become able to make profound generalizations long before they have meet formal algebra in high school. As manager of Mathematics at the body which is now the Victorian Curriculum & Assessment Authority, he was closely involved with the design and implementation of extended assessment tasks in mathematics for the Victorian Certificate of Education. He is a past president of the Australian Association of Mathematics Teachers.

ISDDE, the *International Society for Design and Development in Education*, was formed to bring together from around the world accomplished people who are dedicated to raising the quality of design of educational materials and processes. The goals of the Society are:

- to improve the design and development process;
- to build a coherent professional design and development community;
- to increase the impact of good design on educational practice and policy.

The members are designers and project leaders with outstanding records, together with representatives from government agencies and foundations that fund such work. More on ISDDE and its work can be found at www.isdde.org

[*]1. High-stakes assessment in education systems: roles and challenges

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 learning and teaching that have become accepted internationally, exemplified by the recent Common Core State Standards (CCSS), test validation requires a concomitant rigor with a broad range of strong evidence.

Validity requires tests that are balanced across the performance goals, not just testing those aspects that are easy to test. We were encouraged by the fact that the U.S. Department of Education criteria for award of the assessment consortium grants required the consortia to explain how they would "measure student knowledge and skills against the full range of the college- and career-ready standards, including the standards against which student achievement has traditionally been difficult to measure." Both of the assessment consortia attended carefully to this requirement in their funding proposals (see, e.g., PARCC, p. 54 ff). It is important for both consortia to remain committed to this challenging goal.

The consortia also recognize that they should incorporate an auditing mechanism for checking how well the assessment practice is realizing the intentions (PARCC, p. 208). Such a mechanism should identify problems – for example, unfortunate influences on classroom practice, or deficiencies in assessment and curriculum that leave students unprepared for the higher levels of mathematical thinking and reasoning embodied in the Common Core State Standards. It should then guide appropriate action.

Knowing common patterns of mistakes, and locating student performance in the class along the underlying developmental continua of learning can help teachers plan remediation, as well as change their classroom teaching practices for future students. This formative role for assessment, when well done, can raise standards. Rubrics for scoring and sample responses at each level of performance on the exam can promote changes in classroom practices. Reports that include such responses can be used in discussions with students to provide learning opportunities. Students should see where their performance lies along a progression of learning, so that they 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.

[*] A more general, international version of this paper, without its specific references to PARCC's plans, with be published as part of a forthcoming special issue on assessment design of ISDDE's e-journal *Educational Designer*.

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 per year. Is assessment cost effective in delivering improved education? To make it so is primarily a design and development challenge.

Many of the points made in this paper will be familiar to PARCC, though not necessarily to all those they have to convince to buy into, indeed to buy, high-quality assessment. We have left them in because of their importance in moving systems forward towards better education for all their students.

2. Design principles and practices for summative examinations

In this section we outline the principles that guide the design of tests that aim for high quality. We then describe the practices that enable these principles to be realized. We start with the criteria that articulate what we mean by high quality.

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. To accomplish this, the purposes that the outcomes of the assessment are to be used for, and the likely "backwash" effects on teaching and learning are both important.

A 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, skills and strategies) then to test 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 the educative focus – for example, solving substantial problems.

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

Inferences Can the users[1] 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 standards.

Range and variety Does the variety of tasks in the test match the range of educational and performance aims – as set out in the standards?

Extrapolation Does the breadth and balance of the domain that is assessed justify inferences about the full domain?

The effects the test have 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

[1] 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?". For all users it is important to communicate clearly the performance goals that an assessment is designed to assess.

pattern of learning activities in most of the classrooms tested. 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.

Historically, the validity of narrow tests has sometimes been justified to users as "proxies" for balanced assessment by using evidence of correlation with other measures; but this ignores the effect on classroom practice – and sometimes calls into question the validity of those other measures. Validity of assessments may also be justified by redefining the educational goals to be what the test assesses. These 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, a similar result would be achieved. Reliability is a necessary condition for validity, but not a sufficient one[2]. Some threats to reliability are:

Occasion variability The student may perform at different levels on different days. Performance on an "off day" will give a false impression.

Variability in presentation: The way a task is explained may influence performance in undesired ways, 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 reliably. 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 questions. To narrow the variety of tasks will produce a reduction in validity. Consequently, more extensive and longer assessments will be needed to cover a wider variety of performance types with the same reliability[3]. 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 the tasks is essential to ensure that they remain non-routine; this can offset routine 'teaching to the test' and stereotyping through repetition of the same tasks from year to year, but it may also introduce irrelevant variability.

---

[2] Indeed, some argue that it should be treated as a component of validity. For a more detailed account that gives a full discussion of criteria, all of which are discussed here, see Stobart,G. (2001) The validity of National Curriculum Assessment. *British Journal of Educational Studies* 49 (1) 26-29

[3] 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, and envisioned by PARCC.

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 Language Arts; 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 make 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[4]. Thus, it is important to distinguish between "error", such as mistakes in scoring, and other sources of "variability". These may be either in principle unavoidable or only be avoidable 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; we have noted that some are not even adequate for their declared purpose. Additionally, educational aims do change. For high quality assessment, it is essential that tests can grow along with system improvement.

This is particularly important with new initiatives like PARCC. It is unlikely that current school systems can immediately absorb tests that fully reflect the aims of CCSS; 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 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 – a key educational goal.

removal of weaknesses Tests of higher-level elements of performance present greater design and development challenges that routine tests. They 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, even though the weaknesses could be addressed in other ways[5].

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 contribution of more valid examinations to the quality of education.

Resisting these downward pressures requires an active *engine for improvement* within the assessment system[6]**.** This should combine audit with design research on improving both tests and formative assessment.

---

[4] Qingping He and Dennis Opposs (2010) *A Quantitative Investigation into Public perceptions of Reliability in Examination Results in England.* Coventry : Office of Qualifications and Examinations Regulation. Available for download on : http://www.ofqual.gov.uk/how-we-regulate/133-reliability/415-reliability-programme-technical-advisory-group.

[5] Project work in the classroom presents opportunities for plagiarism or undue teacher "guidance"; these have led to its abandonment rather than to positive moves that improve the assessment procedures.

[6] One obstacle to evolution for improvement can be the aim of measuring whether or not "standards have risen" over time – an aim that requires a consistency over time which, in practice, is probably unattainable. To meet this, old tests can be retained as one component in a more balanced assessment (in PARCC's end of year test, for example). However, controversy in the UK

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 types of task; most current tests draw on too narrow a range.

System design Improvement requires principled changes to assessment systems and their processes; 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 working together:*

Task design The first need is for a pool of tasks, covering the variety of performance goals that the standards imply. Task design involves both systematic and creative skills, so tasks should be collected and/or commissioned 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; and
- 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. The pool of high-quality tasks available worldwide and the design teams that developed them together provide a good start; sustaining this in the longer term will involve a substantial program of designer training.

Task selection and test balancing, under the aegis *of* an authoritative committee – for example, a committee of PARCC, expert in the subject and its assessment. It is in the balancing of the overall assessment that those responsible to the community choose and weight tasks that embody the elements they want to value. This is a crucial role, quite different from creating the range of choice that is the responsibility of broad-spectrum task design; they should be kept independent.

Management of the testing process There are testing agencies that can handle this aspect well. However, this is where most of the cost of testing lies. Competition between test vendors for non-expert customers inevitably favors price over quality. *So for valid tests it is essential that task design and test balancing be independent of vendors.*

Improving the tests is crucial, but it is only a beginning. 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 intra-school and inter-school assessment and monitoring development programs. Such developments require sustained effort but several countries or

---

over such issues led to an investigation by the independent Statistics Commission who concluded that "statutory test data [are] not ideal for monitoring standards over time." Specific research on samples (as in NAEP) offers a better alternative.

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.[7,8]

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, making sure that all components are valued[9].

There are many other sources of evidence about the positive effects of involving classroom teachers in the system. It brings teachers' expertise into assessment. It builds their understanding of the system, and of the educational aims of the examination. This teacher involvement also represents professional development of very high quality. Scoring student responses to complex problems brings out deeper questions about the subject and its values that go beyond assessment, pointing to improvements in teaching. It builds teachers' confidence in the system, while also helping them to form a positive relationship between summative and formative assessment processes and the use of both 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 tasks;
- working actively for improvement at the system level.

We expand on this in Section 5.

Inspiring examples[10]

Nuffield A-level Physics in England[11]included work produced under formal test conditions, and other work produced in more flexible situations. In the former category were:

(1) A multiple-choice test of 40 questions (75 minutes) 20% weighting

[7] Black, P. (2010) *Assessment of and for Learning: improving the quality and achieving a positive interaction.* Invited paper presented to the June 2010 meeting of representatives of the EU education ministers. Brussels: European Union

[8] Hayward, L., Dow, W. and Boyd, B. (2008) Sharing the Standard? Project Report to Scottish Government. Edinburgh: Scottish Education Department.

[9] 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."

[10] There is a fuller description of these and other successful examples in Burkhardt, H. (2009) *On Strategic Design.* Educational Designer, **1**(3). See http://www.educationaldesigner.org/ed/volume1/issue3/article9/index.htm

[11] Pellegrino,J.W., Chudowsky, N. and Glaser, R. (2001) *Knowing what students know: the science and design of educational assessment.* Washington D.C: National Academy Press.Ch.6 p.253-5

(2) A test of 7 or 8 short answer questions (90 minutes) 20%

(3) Questions on a piece of text 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 minutes) 16%

In the second, less formal 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 school-leaving examination was taken by many of the students doing physics in England.

The VCE Mathematics Examination in Australia in the 1990s involved two 90-minute end-of- year exams. Paper 1 was multiple-choice. Paper 2 was extended questions of about 20 to 30 minutes each. During the year there was an investigation project, developed by a panel of experts. For this, due to the high-stakes nature of the assessment, there were concerns about the amount of help students may have received. 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, the external test and internal assessments by schools each contribute 50% to the total scores. Careful studies of inter-correlations between the two scores reveal and explore any anomalies[12].

The novice-apprentice-expert model is one of a number of approaches being developed in the US in response to CCSS. It looks for a balance between:[13]

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 in CCSS (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 task 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. 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.

[12] Stanley, G., MacCann, R., Gardner, J., Reynolds, L. and Wild, I. (2009). *Review of teacher assessment: what works best and issues for development*. Oxford: Oxford University Centre for Educational Development.

[13] Exemplar tasks of each type may be found at http://map.mathshell.org/downloads/map_ccss_ccr_tasks.pdf

The three types are designed to present comparable difficulty, but with a different balance of challenge – largely technical for novice tasks, more strategic and less technical for the others. "Easy expert" tasks, usually allowing a variety of approaches, are at the heart of mathematical literacy[14]

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 in mathematics, rather than just content knowledge, meets resistance, other subjects have shown there is a 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 stake-holders (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, inspiring examples have disappeared with unconnected system changes. This emphasizes the need for *audit* and 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 he or she knows, understands and can do* in the subject.

Scoring 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. 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.) Scoring of essays has much wider variation than scoring short factual answers. The scoring of extended answers is usually carried out by teams of trained scorers. In one common approach, numerical scores are allocated to each part successfully completed. Alternatively, the scoring of rich tasks may use holistic scoring rubrics to construct an overall score based on how well the response meets specific performance criteria. Use of machine scored essays is increasing, but it too may be shown to limit the range of performances that can be assessed. As we stressed in Section 2, improving reliability is a valuable goal only if it is not at the expense of validity.

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

[14] Mathematical literacy is the focus of the Mathematics tests in PISA, the *Programme for International Student Assessment,* organized by OECD, which is now the main instrument for international comparisons of education systems.

variation is usually mentioned, if at all, only in the technical fine print of a test description[15]; 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. *Second scoring* by a senior examiner of all or a sample of papers is common for high-stakes essays. In *consensus monitoring* groups of teachers examine samples of each other's scoring to ensure consistency of standards. We 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[16]. Aggregated scores for classes may provide evidence for teacher evaluation and may be aggregated 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 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 Mathematics tests need to assess performance in mathematics that is balanced across the goals of CCSS.

In summary, different purposes of assessment have a different balance of needs for scoring, monitoring and reporting. Formative and diagnostic assessment in the classroom need detailed feedback. Less and less detail is needed for periodic and final course testing; 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 needed information for the various groups of potential users. 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 make effective scorers of complex tasks on tests[17].

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

[15] For many tests, the test-retest variation is not measured, avoiding any alarm the public might feel if the inherent uncertainty in the results were recognized.

[16] It has been said that "School systems don't buy tests; they buy reports."

[17] See e.g. Foster D., Noyce P. & Spiegel S. *When Assessment Guides Instruction: Silicon Valley's Mathematics Assessment Collaborative* in Schoenfeld, A. H. (Ed.) *Assessing Students' Mathematics Learning: Issues, Costs and Benefits*. Volume XXX, Mathematical Sciences Research Institute Publications. Cambridge, England: Cambridge University Press.

rich tasks are particularly effective, motivating teachers to focus on key issues of performance, content and pedagogy[18].

Communicate the scoring procedures, 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 technology 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 is an essential part of formative assessment. It 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. Technology: its roles, strengths and limitations

There is an understandable enthusiasm in various places for computer-based assessment[19]. It appears to offer inexpensive testing with instant reporting of results. Here we 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 see that technology, 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. Inserting standard responses that check scorer reliability facilitates monitoring. Collecting data for reporting and analysis is universal. Most large-scale test providers use such systems – and, crucially, they present no obvious problems for the wider range of task types that valid assessment of CCSS requires.

[18] Foster D, Poppers A*., Using Formative Assessment to Drive Learning, The Silicon Valley Mathematics Initiative: A Twelve-year Research and Development Project*, Harvard Press 2011 (currently available at www.svmimac.org)

[19] Students who have experienced computer-based tests of mathematics are not always so enthusiastic. (see e.g. Daniel Pead *On computer-based assessment of mathematics,* Nottingham: Shell Centre Publications, or http://www.mathshell.org/papers/dpthesis/)

- *Presenting tasks to students* On-screen presentation is always possible[20]. The potential gain is that it allows a wider variety of tasks. 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. They enable students to explore, analyze, infer and check the properties of a (simulated) system that is new to them.
- *Providing a natural working medium for the student* 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. This is an issue that is often overlooked. 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, 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 for such thinking. Inputting diagrams, fractions or 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 machine scoring more complex responses, major challenges remain. Responses to complex tasks in mathematics and science generally involve sketch diagrams, algebra etc in no particular sequence; there is no proven system for scoring these.

  There is an ongoing danger that the administrative attractions of automatic scoring tempts assessment 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[21]. 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 use in 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)

[20] Research shows that students' responses to on screen and paper versions of the same task are not identical (ibid p.201)

[21] Stacey, K., Price, B., Steinle, V., Chick, H., Gvozdenko, E. (2009) *SMART Assessment for Learning*. Paper presented at the ISDDE Conference in Cairns, Australia. http://www.isdde.org/isdde/cairns/pdf/papers/isdde09_stacey.pdf

- *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 more convenient. 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; they are of little use to teachers. Returning scored papers to students is, unfortunately, less common.

For designers of a high-quality assessment system, the principle is clear: Use technology 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*.

For example, 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 and given 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 cheap and easy proxies for balanced mathematics assessment: multiple-choice and short constructed answer tests with a narrow, fragmented focus, plus some computer-based problem solving tasks with only a distant connection to the mathematical practices in CCSS.

PARCC seems to envision greater use of computer-based testing than the reasoning above would recommend, but without addressing the effects on classroom practice. Experience suggests that these will include:

- teacher belief that computer-based problem solving tasks, rather than the variety of mathematics tasks that CCSS requires, define what mathematical practices are;
- much more computer use in classrooms, but focusing mainly on the new "test prep";
- a widening equity gap, between schools with very different provision and support; and
- huge cost implications.

There could be some positive aspects to this, narrowing the gap between school math and real math, but the result will not be assessing or promoting CCSS. These de-facto changes will raise difficult issues of curriculum and standards, as well as of equity. We conclude this section with a brief discussion of some of them.

The limitations in the usefulness of computers in assessing mathematics is ironical because computers are a central to doing mathematics everywhere outside the classroom, from simple business calculations to research in many subjects, including pure mathematics. But this is not yet reflected in schools where computers and calculators are currently a useful supplement to, not a substitute for, traditional methods in doing or assessing school mathematics. Current curricula and tests mean that most students lack any fluency in the use of spreadsheets[22], computer algebra systems, graphers, dynamic geometry

[22] The predominant use of spreadsheets in schools is for data handling and statistics – typically for producing charts and summatives from survey data – their use as a modelling tool is less common.

packages[23] and (the ultimate mathematical computing tool) programming. These tools would enable students 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 in the taught curriculum to include the practical use of computers in mathematics – a worthwhile end in itself. Students will learn *transferable mathematical* technology 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 fluent with to use during assessments?* We have listed above a range of candidate tools, now used in a minority of schools. (They will still need 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 seems to be 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. It also provides a 'semi-concrete' bridge between arithmetic and the greater abstraction of traditional algebra as a modeling tool.

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, for consideration in the future.

[23] ....including their use for "original" constructions, not just for exploring a pre-prepared model demonstrating a particular concept.

To summarize: given the basic principles that assessment should image desired instruction, because of WYTIWG[24], *and* that desired instruction has to be attainable in ordinary circumstances, a corollary is that computer use in assessment should image desired computer use in instruction.

5. Steering the system

All complex systems depend for their success on the quality of feedback and the mechanisms for using it to guide improvement in system performance[25]. Range, depth and timeliness are all important. Here we link the issues discussed above to the roles of assessment in accountability-based management. In particular, how assessment can be designed and used to guide people at all levels toward realizing the system goals.

Steering versus "Are we 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 "When are we going to get there?". It should be clear that the under-achievements of US education systems, partly reflected by our students' performance in international comparisons, are not going to be solved overnight. Choosing to go for a quick fix guarantees failure. The question is "What *can* be done is to establish directions for change and a program of improvement to move education systems forward with deliberate speed in positive directions?"

Well-designed assessment provides information on direction, not just on distance. In what ways have we improved, as well as 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"[26].

Good assessment provides 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 tests show how far each student has progressed in the various dimensions of performance.

In other areas, assessment provides "canary in the mine" indicators of problem areas 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? To what degree are our professional development programs being delivered as planned? When they are, what changes do we see in teachers' classroom practices? These kinds of question can be answered by evaluative studies, provided these are focused on the goals of the programs concerned.

At 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? Are the pressure points, 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 the leadership is serious and realistic?

---

[24] "What You Test Is What You Get" Burkhardt, H. 1987, *The Dynamics of Curriculum Change* in *Developments in School Mathematics Worldwide*, Wirszup, I & Streit, R. (Eds.) Chicago: University of Chicago School Mathematics Project.

[25] Is education a 'complex adaptive system'? This has been questioned by complexity theorists on the grounds that it does not use even the feedback it collects at classroom level to steer the system, relying on top-down planning based on gross data.

[26] Richard Elmore *Improving The Instructional Core,* Harvard Graduate School of education 2009

At the national level, have we provided school systems with the tools and guidance that will enable their leadership and staff to meet the goals? How well is the national flow of information and support working? Are those in the policy arena engaging teachers and school administrators as decisions of great consequence to schools are made?

Assessment and feedback for teaching and learning,
From the perspective of steering the system we first ask: What should be taught? What should be learned? In CCSS we have a set of "standards" which, with associated exemplification, provide answers that cover the practices of doing mathematics as well as the concepts and skills needed.

The salient 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 no where near sufficient. 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 "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 – not just 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 the capacity to achieve progress within each task not simply between tasks – just as students can improve a piece of writing as they progress, so can they 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. Current tests are not based on growth constructs. They are especially haphazard in their design for the construct of growth in the bottom third, and in 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 and understand; what are their proficiencies? We need work on the design of assessments with tasks to identify the syndromes and detect the growth; currently, 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?* It will be taught to! This needs to be a driver for the quality of the test.
> *Is it worth studying for?* Tests either motivate or demoralize. But they can be educative if the design and use is purposeful.

Multi-dimensional reporting
We have stressed that different users need different kinds and quantities of information. This is not a problem*, provided the needs are all based on a common view of what is important in performance.*

For example, total scores blur together many different dimensions of performance. This may be 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 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, all 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 such consequences 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, a program of regular research on consequences is needed – in particular, information on how teachers, schools, and districts are, and are not, responding to the assessments. This research will tell us where we are steering with the exams. Is it where we want to go? Lack of response or perverse responses may suggest the need for design modifications. Equally, there will be pleasant surprises to build 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 may be a matter of perception, but they are no less important. Past experience has shown that they can be overcome, and the benefits of doing so.

Dangerous 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 the aspects of performance that are actually assessed, as well as their range and balance. This leads 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 per student-test) than machine-scored multiple choice tests ($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 for teachers.
- *Current tests are inexpensive.* While the cost of the test ($1-2) is small, this assertion neglects the great cost of otherwise unproductive "test prep" teaching, which fills many weeks a year in most classrooms – time lost for learning mathematics.
- *Teacher scoring is unreliable*, and subject to cheating. Provided the training and monitoring of teacher scoring are well designed and executed, evidence shows that comparable overall reliability can be achieved.
- *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 the 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, learning and practicing 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.

Resources for moving forward

A wide range of useful resources can be found in a number of countries[27].

Examination systems that embody the various features commended here are, or have been, in general use in various places. There is a body of 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 high quality and that have great potential to "educate", assess and, perhaps of more importance, motivate 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.

The challenges of rethinking large-scale assessments with an eye toward their educative potential for students and teachers seem daunting. But 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 potential that high-quality educative testing practices can become the norm.

7. Conclusion

The aim being pursued in this study is of fundamental importance if assessment is to be designed to encourage the approaches to learning that are needed to prepare students for the future demands of society. The concerns relative to what we need from testing are being articulated and worked on across the world. This is reflected 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)

[27] Too many to list here, ISDDE would be happy to help in pointing them out. Contact Hugh.Burkhardt@nottingham.ac.uk