Literature Review of E-assessment

Jim Ridgway, Sean Mccusker, Daniel Pead

FUTURELAB SERIES

REPORT 10:

Literature Review of E-assessment

Jim Ridgway and Sean McCusker, School of Education, University of Durham
Daniel Pead, School of Education, University of Nottingham

REPORT 10:

Literature Review of E-assessment

Jim Ridgway and Sean McCusker, School of Education, University of Durham
Daniel Pead, School of Education, University of Nottingham

FOREWORD

I have to admit to being someone who for many years has avoided thinking about assessment – it somehow always seemed distant from my interests, divorced from my concerns about how children learn with technologies and, to be honest, just a little less interesting than other things I was working on... In recent years, however, working in the field of education and technology, it has become clear that anyone with an interest in how we create equitable, engaging and relevant education systems needs to think long and hard about assessment. Futurelab's conference 'Beyond the Exam' in November 2003 further highlighted this point, as committed and engaged educators, software and media developers came together to raise a rallying cry for a rethink of our current assessment practices.

What I and many others working in this area have come to realise is that we can't just ignore assessment, or simply see it as 'someone else's job'. Assessment practices shape, possibly more than any other factor, what is taught and how it is taught in schools. At the same time, these assessment practices serve as the focus (perhaps the only focus in this day and age) for a shared societal debate about what we, as a society, think are the core purposes and values of education. If we wish to create an education system that reflects and contributes to the development of our changing world, then we need to ask how we might change assessment practices to achieve this.

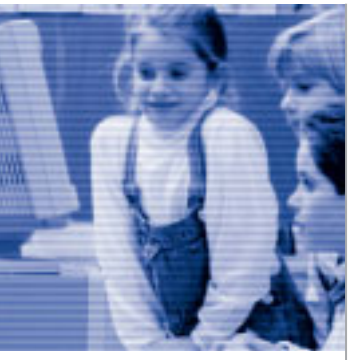
The authors of this review provide a compelling argument for the central role of assessment in shaping educational practice. They outline the challenges and opportunities posed by the changing global world around us, and the potential role of technologies in our assessment practices. Both optimistic and practical, the review summarises existing research and emergent practice, and provides a blueprint for thinking about the risks and potential that awaits us in this area.

We look forward to hearing your response to this review.

Keri Facer, Director of Learning Research
Futurelab
research@futurelab.org.uk

CONTENTS:

EXECUTIVE SUMMARY



assessment is central to educational practice

EXECUTIVE SUMMARY

"E-assessment must not simply invent new technologies which recycle our current ineffective practices."
Martin Ripley, QCA, 2004

Assessment is central to educational practice. High-stakes assessments exemplify curriculum ambitions, define what is worth knowing, and drive classroom practices. It is essential to develop systems for assessment which reflect our core educational goals, and which reward students for developing skills and attributes which will be of long-term benefit to them and to society. There is good research evidence to show that well designed assessment systems lead to improved student performance. In contrast, the USA provides some spectacular examples of systems where narrowly focused high-stakes assessment systems produce illusory student gains; this 'friendly fire' results at best in lost opportunities, and at worst in damaged students, teachers and communities.

ICT provides a link between learning, teaching and assessment. In school, ICT is used to support learning. Currently, we have bizarre assessment practices where students use ICT tools such as word processors and graphics calculators as an integral part of learning, and are then restricted to paper and pencil when their 'knowledge' is assessed.

Assessment systems drive education, but are themselves driven by a number of factors, which sometimes are in conflict. To understand likely developments in assessment, we need to examine some of these drivers of change. Implications of technology, globalisation, the EU, multinational companies, and the need to defend democracy are discussed. All of these influences are drivers for increased uses of ICT in assessment. Many of the developments require the assessment of higher-order thinking. However, there is a constant danger that assessment systems are driven in undesirable ways, where things that are easy to measure are valued more highly than things that are more important to learn (but harder to assess). In order to satisfy educational goals, we need to develop ways to make important things easier to measure - and ICT can help.

All is not well with education. The Tomlinson Report (2004) identifies major problems with current educational provision at ages 14-19 years: there is a plethora of qualifications; too few students engage with education; the drop-out rate is scandalously high; and the most able students are not stretched by their studies. Young people are not being equipped with the generic skills, knowledge and personal attributes they will need in the future. A radical approach to qualifications is suggested which (in our view) can only be introduced if there is a widespread adoption of e-assessment.

The UK government is committed to a bold e-assessment strategy. Components include: ICT support for current paper-based assessment systems; some online, on-demand testing; and the development of radical, ICT-set and assessed tests of ICT capability. Some good progress has been made with these developments.

E-assessment can be justified in a number of ways. It can help avoid the meltdown of current paper-based systems; it can assess valuable life skills; it can be better

for users – for example by providing on-demand tests with immediate feedback, and perhaps diagnostic feedback, and more accurate results via adaptive testing; it can help improve the technical quality of tests by improving the reliability of scoring.

E-assessment can support current educational goals. Paper and pencil tests can be made more authentic by allowing students to word process essays, or to use spreadsheets, calculators or computer algebra systems in paper-based examinations. It can support current UK examination processes by using Electronic Data Exchange to smooth communications between schools and examinations authorities; current processes of training markers and recording scores can be improved. Systems where student work is scanned then distributed have advantages over conventional systems in terms of logistics (posting and tracking large volumes of paper, for example), and continuous monitoring can ensure high marker reliability. Current work is pushing boundaries in areas such as text comprehension, and automated analysis of student processes and strategies.

E-assessment can be used to assess 'new' educational goals. Interactive displays which show changes in variables over time, microworlds and simulations, interfaces that present complex data in ways that are easy to control, all facilitate the assessment of problem-solving and process skills such as understanding and representing problems, controlling variables, generating and testing hypotheses, and finding rules and relationships. ICT facilitates new representations, which can be powerful aids to learning. Little is known about the cognitive implications of these representations; however, it seems likely that complex ideas (notably in reasoning from evidence of various sorts) will be acquired better and earlier than they are at present, and that the standards of performance demanded of students will rise dramatically. Here, we also explore ways to assess important but ill-defined goals such as the development of metacognitive skills, creativity, communication skills, and the ability to work productively in groups.

A major problem with education policy and practice in England is the separation of 'academic' and 'practical' subjects. In the worst case, to be able to invent and create something of value is taken to be a sure sign of feeble-mindedness; where as to opine on the work of others shows towering intellectual power. A diet of academic subjects with no opportunities to act upon the world fails to equip students with ways to deal with their environments; a diet of practical subjects which do not engage higher-order thinking throughout the creative process equip students only to become workers for others. Both streams produce one-handed people, and polarised societies. E-portfolios can provide working environments and assessment frameworks which support project-based work across the curriculum, and can offer an escape from one of the most pernicious historical legacies in education. E-portfolios solve problems of storing student work, and make the activity of documenting the process of creation and reflection relatively easy. Reliable teacher assessment is enabled. There is likely to be extensive use of teacher assessment of those aspects of performance best judged by humans (including extended pieces of work assembled into portfolios), and more extensive use made of on-demand tests

e-assessment can be used to assess 'new' educational goals

e-assessment is a stimulus for rethinking the whole curriculum

of those aspects of performance which can be done easily by computer, or which are done best by computer.

The issue for e-assessment is not **if** it will happen, but rather, **what**, **when** and **how** it will happen. E-assessment is a stimulus for rethinking the whole curriculum, as well as all current assessment systems. New educational goals continue to emerge, and the process of critical reflection on what is important to learn, and how this might be assessed authentically, needs to be institutionalised into curriculum planning.

E-assessment is certain to play a major role in defining and implementing curriculum change in the UK. There is a strong government commitment to high quality e-assessment, and good initial progress has been made; nevertheless, there is a need to be vigilant that the design of assessment systems is not driven by considerations of cost.

Major challenges of 'going to scale' have yet to be faced. A good deal of innovative work is needed, coupled with a grounded approach to system-wide implementation.

PURPOSE

The purpose of this report is:

- to assert the centrality of assessment in education systems
- to identify 'drivers' of assessment, and their likely impact on assessment, and thence on education systems
- to describe current, radical plans for increased use of high-stakes e-assessment in the UK
- to describe and exemplify current uses of ICT in assessment
- to explore the potential of new technologies for enhancing current assessment (and pedagogic) practices
- to identify opportunities and to suggest ways forward
- to 'drip feed' criteria for good assessment throughout (set out explicitly in an appendix).

This report has been designed to: present key findings on research in assessment; describe current UK government plans, and likely future developments; provide links to interesting examples of e-assessment; offer speculations on possible future developments; and to stimulate a debate on the role of e-assessment in assessment, teaching, and learning.

The key findings and implications of the report are presented within the Executive Summary.

ASSESSMENT DRIVES EDUCATION

1 ASSESSMENT DRIVES EDUCATION

Assessment is an integral part of being. We all make myriads of assessments in the course of everyday life. Is Jane a good friend? Which Rachel Whiteread do I like best? Does my bum look big in this? The questions we ask, and the referents, give an insight into the way we see ourselves and the world (eg Groucho Marx's "Please accept my resignation. I don't want to belong to any club that will accept me as a member"). For aspects of our lives that are goal-directed (getting promoted, going shopping), assessment is essential to progress. To be effective, it is necessary to know something of the intended goal; in well-defined situations, this will be relatively easy, and goals will be specified clearly. In ill-defined situations, such as creative acts, and research, the goals themselves might not be well specified, but the criteria for assessing products and processes may well be.

1.1 ASSESSMENT AND EDUCATION

Assessment is central to the practice of education. For students, good performance on 'high-stakes' assessment gives access to further educational opportunities and employment. For teachers and schools, it provides evidence of success as individuals and organisations. Cultures of accountability drive everyone to be 'instrumental' – how do I demonstrate success (without compromising my deep values)? Assessment systems provide the ways to measure individual and organisational success, and so can have a profound driving influence on systems they were designed to serve.

There is an intimate association between teaching, learning and assessment, illustrated in Fig 1. Robitaille et al (1993) distinguish three components of the curriculum: the intended curriculum (set out in policy statements), the implemented curriculum (which can only be known by studying classroom practices) and the attained curriculum (which is what students can do at the end of a course of study). The links between these three aspects of the curriculum are not straightforward. The 'top down' ambitions of some policy makers are hostages to a number of other factors. The assessment system – tests and scoring guides - provides a far clearer definition of what is to be learned than does any verbal description (and perhaps provides the only clear definition), and so is a far better basis for curriculum planning at classroom level than are grand statements of educational ambitions. Teachers' values and competences also mediate policy and attainment; however, the assessment system is the most potent driver of classroom practice.

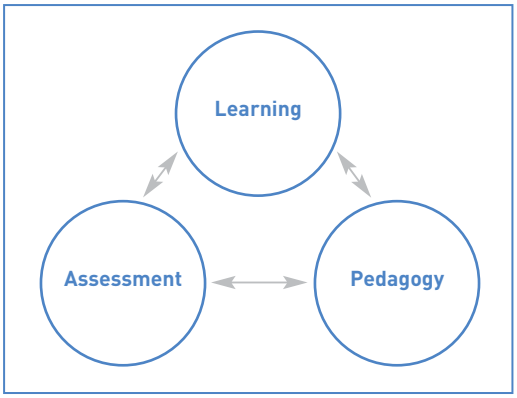


Fig 1: Adapted from Pellegrino, Chudowski and Glaser (2001)

the assessment system is the most potent driver of classroom practice

ASSESSMENT DRIVES EDUCATION

In the UK, there is a long-standing belief (eg Cockcroft 1982) that assessment systems have a direct effect on curriculum and on classroom practices. In Australia, Barnes, Clarke and Stevens (2000) traced the effects of changing a high-stakes assessment on classroom practice, and claimed evidence for a direct causal link. Mathews (1985) traced the distorting effects on the whole school curriculum of formal examinations for university entrance (now A-levels), introduced when the university sector expanded beyond Cambridge, Durham and Oxford – to accommodate as much as 5% of the population. There was a perceived need for entrance tests to pre-university courses (O-levels) – designed for about 20% of the population - followed by a perceived need to align all certification in the education system (notably O-levels and CSE). This linkage between assessment for university admission and the assessment of low-attaining students had a direct and often damaging impact on courses of study for lower attaining students (Cockcroft 1982).

Ill-conceived assessment can damage educational systems. Klein, Hamilton, McCaffrey and Stecher (2000) present evidence on the 'Texas Miracle'. Here, scores on a rather narrow test designed by the State of Texas showed very large gains over a period of just four years. This test is used to determine the funding received by individual schools. Unfortunately, scores on a national test which supposedly measured the same sort of student attainment were largely unchanged in the same time interval. So scores on narrow tests can rise, even when underlying student attainment does not. The 'Texas Miracle' was used in the election campaign of President Bush, as evidence of his effectiveness as a governor in raising educational standards.

ill-conceived assessment can damage educational systems

Linn (2000) points to an underhand method sometimes used by incoming superintendents of school districts to show the effectiveness of their leadership. Most commercially available multiple choice tests of educational attainment have a number of 'parallel test forms', designed to measure the same knowledge and skills in the same way, but with slightly different formats (so '12 men take six days, how long will six men take?' becomes '12 men take six days, how long will four men take?'). These tests are designed in such a way that student scores on two parallel forms would be the same (plus or minus measurement error). Test designers do this so that school districts can change the test form every year, in order that tests measure the underlying knowledge and skills, not the ability to memorise the answers to specific questions. Linn (2000) gives an example where an incoming Superintendent decides to use a new test form and also chooses to use this same test form in successive years. The result is a steady increase in student scores simply because of poor test security – students are taught to memorise answers. It appears that the superintendent has worked miracles with student attainment, because scores have gone up so much. However, when students are tested on a new parallel form, and have to work out the answers and not rely on memory, then scores plummet. So the high reputation for increasing student performance is built upon deliberate deceit. This is bad for teachers and students, and bad for public morality.

High-stakes assessment systems define what is rewarded by a culture, and

therefore the knowledge that is valuable. It is unsurprising that high-stakes assessment has a profound effect on both learning and teaching. Decisions about assessment systems are not made in a vacuum; the educational community in the UK (but not universally) is involved in the design of assessment systems, and these decisions are usually grounded in discussions on what is worth knowing, and in the practicalities of teaching different concepts and techniques to students of different ages.

1.2 THE IMPACT OF ASSESSMENT ON ATTAINMENT

An extensive literature review by Black and Wiliam (2002) showed that well designed formative assessment is associated with major gains in student attainment on a wide range of conventional measures of attainment. This result was found across all ages and all subject disciplines. Topping (1998) reviewed the impact of peer assessment between students in higher education on writing, and found large positive effects. A major literature review commissioned by the EPPI Centre (2002) showed that regular summative assessment had a large negative effect on the attainment of low-attaining students, but did little harm to high-attaining students. These studies provide strong evidence that good assessment practices produce large performance gains. These gains are amongst the largest gains found in any educational 'treatments'. Similarly, poor assessment systems have negative – not neutral – effects on the performance of weak students. It follows that when we consider the introduction of e-assessment, we should be aware that we are working with a very sharp sword.

1.3 ICT AND ASSESSMENT

ICT perturbs the links between learning, teaching and assessment in a number of distinct ways:

1 ICT has changed the ways that research is conducted in most disciplines. Linguists analyse large corpuses of text; geographers use GIS systems; scientists and engineers use modelling packages. Everyone uses word processors, databases and spreadsheets. Students should use contemporary research methods; if they do not, school-based learning will become increasingly irrelevant to understanding developments in knowledge. Assessment should reinforce good curriculum practice. We are approaching a bizarre situation where students use powerful and appropriate tools to support learning and solve problems in class, but are then denied access to these tools when their 'knowledge' is assessed.

2 ICT can support educational goals that have been judged to be desirable for a long time, but hard to achieve via conventional teaching methods. In particular, ICT can support the development of higher-order thinking skills such as critiquing, reflection on cognitive processes, and 'learning to learn', and can facilitate group work, and engagement with extended projects; ICT competence is itself a (moving) target for assessment.

3 New technologies raise an important set of questions about what is worth learning in an ICT-rich environment; what can be taught, given new pedagogic tools; and how assessment

well designed formative assessment is associated with major gains in student attainment

ASSESSMENT DRIVES EDUCATION

systems can be designed which put pressure on educational systems to help students achieve these new goals. If we ignore these important questions, we run the risk that e-assessment will be designed on the basis of convenience, with disastrous consequences for educational practice.

1.4 ON THE NATURE OF SUMMATIVE AND FORMATIVE ASSESSMENT

We should distinguish between summative and formative assessment, which are different in conception and function. In principle, it is easy to distinguish between them. Summative assessment takes place at the end of some course of study, and is designed to summarise performance and attainment at the time of testing; high-stakes, end of schooling assessment such as GCSE provides a good example. Formative assessment takes place in mid-course, and is intended to enhance students' final performance; comments on the first draft of an essay provide an example.

Summative and formative assessments differ on a number of dimensions. These include:

Consequences: summative assessment is often highly significant for the student and teacher, whereas formative assessments need not be.

Exchange value: summative assessments often have a value outside the classroom - for certification, access to further courses, and careers; formative assessment usually has no currency outside a small group.

Audience: summative evaluations often have a large audience; the student and teacher, parent, school, employer and educational system. Formative evaluation can have a small audience; perhaps just the student and teacher (and parent in younger years).

Mendacity quotient: in summative assessment, students are advised to focus on things they do best and hide areas of ignorance; in formative assessment, it is more sensible for students to focus on things they understand least well.

Agency: summative assessment is often done to students, perhaps without their willing participation. Formative assessment is often actively sought out by the student; good formative feedback depends on student engagement in the process of revision.

Validation methods: summative assessment is often judged in terms of predictive validity - are students who got A grades more likely to get top grades in college (but see Messick 1995)?? Formative assessment might be judged in terms of its usefulness in undoing predictive validity – what feedback can we give to students with C grades, so that they perform as well in college as anyone else?

Quality of the assessment: for summative assessment, the assessment method should achieve appropriately high standards of reliability and validity; for formative assessment, 'reliability and validity' are negotiable between teacher and student.

Resources required: the nature of summative assessment can be influenced by considerations of cost and time. In

terms of cost, the estimation of the cost of testing is often done very badly, especially in the USA. There, it is common for 'cost' to be equated with the money paid for the test and its scoring, not the real cost, which is the opportunity cost, measured in terms of the reduction in time spent learning which has been diverted to useless 'test prep'. Formative evaluation should be an integral part of the work of teaching, so estimation of cost focuses naturally on opportunity costs – just what is an effective allocation of teaching and learning time to formative evaluation? In terms of time, for summative assessment time is easy to measure (so long as useless 'test prep' is counted in); again, formative assessment is an integral part of teaching.

Knowledge and the knowledge community: summative assessment is explicit about what is being assessed, and ideas about the nature of knowledge are shared within a wide community; with formative evaluation, ideas about the nature of knowledge might be negotiated by just two people.

Status of the assessment: in summative assessment, the assessment can be ignored by the student; formative assessment simply isn't formative assessment unless the student does something with it to improve performance.

Focal domain: it is useful to distinguish between cognitive, social and emotional aspects of performance. Summative assessment commonly focuses on cognitive performance; formative assessment can run wild in the social and affective domains.

Theory dependence: summative assessment rarely rests on theory; formative assessment is likely to be 'theory-genic' as participants discuss progress, what is known, how to learn and remember things, and how best to use evidence.

Tool types: summative assessment commonly uses timed written assessments where the structure is specified in advance, and which is scored using a common set of rules. Tests are often designed to discriminate between students, and to put them into a rank order in terms of performance. Formative assessment commonly uses a variety of methods such as portfolios of work, student draft work, student annotations of their work, concept mapping tools, diagnostic interviews and diagnostic tests. Each student is their own referent – comparison with other students may not be useful, and is often harmful to learning.

1.4.1 Reflecting on summative and formative assessment

Despite the differences highlighted here, the two sorts of assessment have many areas of overlap:

- a student can change their study methods on the basis of an end-of-year examination result (summative assessment used for formative purposes)
- summative evaluation of students can provide formative evaluation for teachers, schools and educational systems
- formative assessment always rests on some sort of summative assessment – feedback and discussion must rest

ASSESSMENT DRIVES EDUCATION

frequent testing and reporting of scores damages weaker students

on some assessment of the current state of knowledge

- some summative assessment should include the ability to benefit from formative assessment – learning to learn is an important educational goal, and should be assessed, formally
- summative assessment (eg of student teachers) should include the ability to provide formative assessment.

1.5 SUMMARY OF SECTION 1

Assessment lies at the heart of education. Assessment systems exemplify the goals and values of education systems. High-stakes assessment systems have a direct influence on classroom practices. Any discussion of assessment raises important questions about what is worth knowing, the extent to which such knowledge can be taught, and the best ways to support knowledge acquisition.

Well designed assessment systems are associated with large increases in student performances; frequent testing and reporting of scores damages weaker students. Badly designed high-stakes assessment systems can have strong negative consequences for students, communities and societies.

In this section, we distinguish between summative assessment (assessment of learning) and formative assessment (assessment for learning), and compare their characteristics.

ICT has changed the ways that academic work is done; this should be reflected in the tools used in education for both learning and assessment. Bizarre current practices where ICT is an integral part of learning, but where students are denied access to technology during assessment, must be reformed as a matter of urgency. Skills in ICT are essential for much modern living, and so should be a target for assessment.

HOW AND WHERE MIGHT ASSESSMENT BE DRIVEN?

2 HOW AND WHERE MIGHT ASSESSMENT BE DRIVEN?

There is a comforting belief that decisions about education and education systems are made within those systems, and that outside agents – notably foreign outside agents – have little or no influence on internal affairs. This has been true in the UK for a long time, but has not been true in countries which (for example) make use of UK examinations to certify students. If we are to explore plausible scenarios about the future impact of ICT on assessment, it is necessary to take account of 'drivers of change'. Here, we consider technology, globalisation, the rise of mass education, problems of political stability, current government plans, and likely government plans, as drivers of educational change and, in parallel, of likely changes in assessment systems.

2.1 TECHNOLOGY AS DRIVER OF SOCIAL CHANGE

Technology is a key driver of social change. Technology has transformed the ways we work, our leisure activities, and the ways we interact with each other. The use of the web is growing at an extraordinary rate, and people increasingly have access to rich sources of information. Metcalfe's law states that the value of a network rises dramatically as more people join in – its value doesn't just increase steadily. The capability of computer hardware and software continues to improve, and features are being added (such as high quality video) which make computer use increasingly attractive, and well suited to supporting human-human interactions. The web is an increasingly valuable resource which is becoming progressively easier to use, and is attracting users at an increasing rate. Technology is ubiquitous: as well as computers in the form of desktops and laptops, there has been an explosion of distributed computer power in the form of mobile phones which are also fully functioning personal digital assistants (PDAs), containing features such as a spreadsheet, database and word processor. It has been estimated that there are over three billion mobile phones worldwide (Bennett 2002); as before, this number is growing very fast, and new phones are manufactured with an increasing range of features. Technology as a driver has a number of likely effects on assessment. New skills (and so new assessments) are needed for work and social functioning, which require fluent use of ICT; technology has had a profound effect on many labour intensive work practices, many of which resemble educational assessment. The use of ICT for assessment has hardly begun, and some new technologies such as mobile phones offer great promise not only because of their ubiquity (which might solve a current problem of access which has restricted widespread use of ICT in assessment in the past), but also because new technologies have become a natural form of communication for very many young people.

2.2 GLOBALISATION

Globalisation is probably the most obvious driver of change. Significant features for the current discussion are: the mobility of capital, employment opportunities (jobs), and people. Cooperation between countries (eg in the European Union), and the pervasive influence of multinational companies also have profound social effects.

the use of ICT for assessment has hardly begun

HOW AND WHERE MIGHT ASSESSMENT BE DRIVEN?

cooperation between countries in Europe will have an effect on assessment systems

The mobility of capital and jobs has changed the profile of the job market, with new kinds of jobs being created (eg in ICT) and old ones disappearing (eg in manufacturing industries). It is very easy to export jobs and capital from the developed world to the developing world (eg by relocating telephone call centres, or by establishing factories in countries with low wage costs). For people (and economies) to be successful, they must continue to learn new skills, and to adapt to change. Retraining will often require re-certification of competence, with the obvious consequence of further assessment, and the need to design assessment systems appropriate to the new needs of employment. These are pressures for more, and effective, systems of competence-based assessment.

Migration for work and education raises similar issues. The developed world has a need to import highly skilled workers; universities worldwide seek international students. In both cases, there is a need to certify the competence of applicants, and to reject those least likely to be effective workers, or to complete courses successfully (because of a lack of fluency in the language of instruction, for example). Financial considerations make it impractical for testing to take place in the target country, and so a good deal of testing takes place in the country supplying workers or students. Again, it is common to use competence tests which are externally mandated and designed. Language testing provides a good example; a computer-based version of the Test of English as a Foreign Language (TOEFL) has been developed which adjusts the difficulty level of the questions in the light of the performance of the candidate on the test (see www.ets.org/toefl).

For developed economies to maintain their global dominance, their economies must be geared to 'adding value' to raw materials (or to creating value from nothing, as in the entertainment and finance industries). This requires changes in the education system which encourage creative activities, and good problem-solving ability. Employment in a post-industrial society is likely to depend on higher-order thinking skills, such as 'learning to learn'. This requires that these thinking skills be exemplified and assessed, if they are to receive appropriate attention in school.

The effects of cooperation between countries in Europe will have an effect on assessment systems. Currently, there is a problem that qualifications in different member states ('architect', 'engineer') are gained after rather different amounts of training, and equip people for quite different levels of professional responsibility. This makes job mobility very difficult. The Bologna Accord is an agreement between EU member states that all universities will adopt the same pattern of professional training (typically a three-year undergraduate degree followed by a two-year professional qualification) in order to make qualifications in different member states more comparable. Convergence of course structure is likely to lead to a convergence of assessment systems, in line with the desire to increase mobility (see www.engc.org.uk/international/bologna.asp for an analysis of the impact of the Bologna, Washington and Sidney Accords on engineering).

Globalisation is having a profound effect on educational systems worldwide. In higher education, Slaughter and Leslie (1997) describe the response of universities in

several countries to 'academic capitalism' – a global trend to view knowledge as a 'product' to be created and controlled, and to see universities as organisations which produce knowledge and more knowledgeable people as efficiently as possible. They document the changes in university structures and functioning which have been a response to such pressures; these include greater collaboration on teaching between universities, and mutual accreditation of courses. Again, the need for comparability of course difficulty and student attainment will lead to a careful re-examination of assessment systems, and some homogenisation.

Multinational companies also drive changes in assessment practices. These companies are successful in part because of their emphasis on uniform standards; one is unlikely to get a badly cooked hamburger in Macdonalds, or a copy of Excel that functions worse than other copies. This emphasis on quality control extends to job qualifications, and to standards required of workers. In fast changing markets such as technology provision, retraining workers and checking their competence to use, install or repair new equipment or software requires appropriate assessment of competence. The needs of employers for large numbers of staff who are able to use ICT effectively as part of their job has lead to trans-national qualifications such as the European Computer Driving Licence (www.ecdl.co.uk). Such examples are interesting because they are set by international organisations, or commercial organisations, and in some cases (eg the Microsoft Academy programme - www.microsoft.com/education/msitacademy/ITAPApplyOnline.aspx), state-funded educational organisations must submit themselves for examination by a commercial company before they are allowed to certify student competence.

The scale on which such examinations are taken is impressive. Bennett (2002) describes the National Computer Rank Examination, China, which is a proficiency exam to assess knowledge of computer science and the ability to use it; two million examinations were taken in 2002. Tests for the European Computer Driving Licence have been taken by more than a million people.

2.3 MASS EDUCATION

Mass education has developed rapidly and recently. In the last 30 years, the percentage of the UK population being educated at university has risen from about 5% to about 40%. This puts pressures on academic systems to develop efficient assessment systems.

There is now a great deal of distance education. China plans to have five million students in 50-100 online colleges by 2005. At least 35 US states have virtual universities (Bennett 2002). (The recent failure of the E-university in the UK - www.parliament.uk/post/pn200.pdf - and of the US Open University, shows that such ventures are not always successful!) A great deal of curriculum material is delivered via a variety of technologies (the Massachusetts Institute of Technology is in the process of putting all its course material online, for example – see http://ocw.mit.edu/index.html). Over 3,000 textbooks are freely available online at the National Academy Press (www.nap.edu). The use of technology in the assessment process is a logical consequence of these developments.

multinational companies also drive changes in assessment practices

HOW AND WHERE MIGHT ASSESSMENT BE DRIVEN?

2.4 DEFENDING DEMOCRACY

Problems of potential political instability provide another driver of change. The rise of fundamentalism (both Christian and Moslem) can be seen as a loss for rationalism. Electoral apathy is a threat to the democratic process. One problem for politicians is to explain complex policies to citizens. This is made difficult if citizens understand little about modelling (such as ideas of multiple causality, feedback in systems, lead and lag times of effects etc). Informed citizens need to understand something about ways to describe and model complex systems, in order that they do not give up on democracy simply because they do not understand the policy arguments being made. Understanding arguments about causality and some experience of modelling systems via ICT should be major educational goals. These goals will need to be exemplified and valued by high-stakes assessment systems, if they are to become part of students' educational experiences.

Education for citizenship has received increasing emphasis in the UK. Some of the educational goals – such as understanding different perspectives, increased empathy, and community engagement - seem intangible. However, ICT can play a role in posing authentic questions (for example via video) and could play a role in formative assessment, and perhaps in summative assessment (using portfolios).

continued expansion of the current examination system without some changes does not seem a viable option

2.5 GOVERNMENT-LED REFORMS IN CURRICULUM AND ASSESSMENT

Governments are responsive to global pressures, and analyses of the limitations of current national systems. Two current UK initiatives are likely to lead to radical changes in assessment practices, notably to increase the use of e-assessment. One is the DfES E-assessment Strategy (www.dfes.gov.uk/elearningstrategy/default.stm) which maps out a tight timeline for change in current examination systems; the other is the Tomlinson (2004) Report 14-19 Curriculum And Qualifications Reform, which proposes radical changes in educational provision itself (with direct consequences for e-assessment).

The Tomlinson Report (2002) into A-level standards argued that the examinations system is operating at, or perhaps beyond, capacity. According to Tomlinson (2002), in 2001, 24 million examination scripts and coursework assignments were produced at GCSE, AS and A level. In terms of the number of students being assessed, in 2002 there were around six million GCSE entries and nearly two million children sat Key Stage tests. More students are engaging in post-compulsory education; the introduction of modular A-levels, and the popularity of AS courses has resulted in an increase in the number of examinations taken (Tomlinson reports a growth of 158% over a 20-year period). There is an associated problem concerning the supply of examiners, in terms of both recruitment and training. Roan (2003) estimated that about 50,000 examiners were involved in the assessment of GCSEs, GNVQs and A-levels. Continued expansion of the current examination system without some changes does not seem a viable option. ICT support for current activities, described later, might well be of benefit.

ICT-based assessment is now part of UK government policy, and will be introduced

progressively, but on a tight timescale. The DfES E-learning Strategy will be accompanied by radical changes to the assessment process, for which the Qualifications and Curriculum Authority are responsible (www.qca.org.uk/adultlearning/workforce/6877.html). Over the next five years, the following activities are planned:

"All new qualifications should include assessment on-screen

Awarding bodies set up to accept and assess e-portfolios

Most examinations should be available optionally on-screen, where appropriate

National curriculum tests available on-screen for those schools that want to use them

The first on-demand GCSE examinations are starting to be introduced

10 new qualifications specifically designed for electronic delivery and assessment"

QCA Blueprint (2004)

The timescale for these changes is short. For example, in 2005, 75% of basic and key skills tests will be delivered on-screen; in 2006, each major examination board will offer live GCSE examinations in two subjects, and will pilot at least one qualification, specifically designed for electronic delivery and assessment; in 2007, 10% of GCSE examinations will be administered on-screen; in 2008, there will be on-demand testing for GCSEs in at least two subjects.

Good progress has been made with these developments. For example, Edexcel is carrying out a pilot scheme for online GCSEs in chemistry, biology, physics and geography with 200 schools and colleges across the West Midlands and the west of England. AQA conducted a live trial in March 2004 on 20,000 scripts (Adams and Hudson 2004); in Summer 2004, about 500,000 marks (5% of the total) will be collected; by 2007, 100% of marks will be captured electronically.

The Tomlinson Report (2004, in prep) will offer a more radical challenge to assessment practices. The Interim Report (Tomlinson 2004) identified a number of problems with the existing system. These include concerns about:

- excellence – the current system does not stretch the most able young people (in 2003, over 20% of A-level entries resulted in grade A)
- vocational training – there is an historic failure to provide high-quality vocational courses that stretch young people and prepare them for work
- vocational learning is often assessed by external written examinations, not practical and continuous assessment
- assessment - the burden on students and teachers is too high
- disaffection - our high drop-out rates are scandalous
- the plethora of qualifications – currently around 4,000
- curricula - are often narrow, overfull, and limit in-depth learning
- too few students develop high levels of competence in mathematical skills, communication, working with others, or problem-solving
- failure to equip young people with the generic skills, knowledge and personal attributes they will need in the future.

the Tomlinson Report will offer a more radical challenge to assessment practices

HOW AND WHERE MIGHT ASSESSMENT BE DRIVEN?

there is an urgent need to invent and apply new sorts of e-assessment on a large scale

The Report proposes a single qualifications framework, based on diplomas set at four levels (Entry, Foundation, Intermediate and Advanced). Students are expected to progress at a pace appropriate to their attainment, rather than their age. Each diploma shares some common features. These require students to demonstrate evidence of:

- mathematical skills, communication and ICT skills
- successful completion of an extended project
- participation in activities based on personal interest, contribution to the community as active citizens, and experience of employment
- personal planning, review and making informed choices
- engagement in 'main learning'- the major part of the diploma – chosen by the student in order to open access to further opportunities (eg in employment or education).

These recommendations are exciting and very ambitious, but deeply problematic, unless there are radical changes to current assessment systems – notably in the large-scale adoption of e-assessment. We consider ways these recommendations might be met, in Section 3.

2.6 SUMMARY OF SECTION 2

A number of 'drivers' are shaping both assessment and ICT; these need to be taken into account in any discussion of future developments. These drivers provide conflicting pressures. The drivers considered here include the increasing power and ubiquity of ICT, and the explosion of its usefulness and use in everyday life. These provide pressures for more relevant skills to be assessed, and also provide an assessment medium which is largely unexplored. Demands for lifelong learning, for people who can innovate and create new ideas, and the needs for informed citizenship are all pressures for education (and associated assessment systems) that rewards higher-order thinking, and personal development. Conversely, drivers such as the need to retrain and recertify staff, to ensure common standards across organisations in different countries, and to allow access to well-qualified migrants for jobs and education, emphasise assessments which transcend national boundaries and which are based on well-defined competencies (and where assessment design is sometimes based on perceived commercial imperatives). These drivers require different approaches to assessment, and all require new sorts of assessments and assessment systems to be developed.

In the UK, there are a number of problems with current assessment systems. First, they serve students very badly; second, they might soon collapse under their own weight. There is now the political will (and a tight timescale) to develop pervasive, high quality e-assessment on a tight timeline, aligned with current and emerging educational goals. There is also an urgent need to invent and apply new sorts of e-assessment on a large scale.

CURRENT DEVELOPMENTS IN E-ASSESSMENT

3 CURRENT DEVELOPMENTS IN E-ASSESSMENT

The UK government has embarked on a very ambitious project to extend the use of e-assessment. The issue for education is not if e-assessment will play a major role, but when, what, and how. E-assessment can take a number of forms, including automating administrative procedures; digitising paper-based systems, and online testing - which extends from banal multiple choice tests to interactive assessments of problem-solving skills. In this section, we focus on current developments in e-assessment for summative purposes that can be used across the educational system. In Section 4 we address important but less well-defined targets for e-assessment.

Before we begin this section exploring different aspects of e-assessment, we should remember some of the virtues of paper-based tests, in order that we do not become so enamoured of new technologies that we lose sight of the benefits of current assessment systems. With paper:

- all stakeholders are familiar with all aspects of the medium
- paper is robust – it can be dropped, and it still functions
- there are rarely problems of legibility
- high resolution displays are readily available
- students can take questions in any order
- users can input cursive script, diagrams, graphs, tables
- a number of equity issues have been solved – it is easy to create large fonts and to solve other access problems
- paper-based testing systems are well established - it is relatively easy to prevent candidates from copying from each other, for example
- paper is easy to distribute, and can be used in most locations
- in extreme circumstances, it is possible to copy an examination paper, and find another desk
- human judgements are brought to bear throughout the process, so the scope of questions is unconstrained.

3.1 SOME MOTIVES FOR COMPUTER-BASED TESTING

A number of justifications have been put forward for computer-based testing, and are set out below. Not all justifications apply to every use of computers in assessment.

Avoiding meltdown: it may well be impossible to maintain existing paper-based assessment systems in the face of the current growth in the number of students being tested. Scanning technologies can help.

Valuable life skills: much of everyday life (including professional life) requires people to use computers. Not using computers for assessment seems perverse.

Alignment of curriculum and assessment: there is a danger of an emerging gap between classroom practices and the assessment system. It is very common for students (and almost all professionals) to use word processors when they write; in mathematics and science, the use of graphics calculators, spreadsheets, computer algebra systems (CAS) and

the issue for education is not if e-assessment will play a major role, but when, what, and how

CURRENT DEVELOPMENTS IN E-ASSESSMENT

on-demand testing would enable students to take tests when they are ready

modelling software is commonplace (and universal in professional practice). Assessment systems that do not allow access to these tools are requiring students to work in unfamiliar and maladaptive ways. Non-ICT-based assessment can be a drag on curriculum reform, rather than a useful driver (see Section 1.2).

On-demand testing: in many situations (for example, students engaged in part-time study; students taking courses designed to develop competencies; students on short courses) it is appropriate to test students whenever they are judged (or judge themselves) to be ready. City and Guilds tests provide an illustration; 75,000 online tests have been taken, and candidates book a test time that suits them. Saturday is the third most popular day for assessment (Ripley 2004).

Students progress at different rates: currently, the UK examination system acts as a force against differentiation in the curriculum. Summative end-of-year tests make it attractive to schools to teach year groups together and to enter them in a common set of examinations. On-demand testing would enable students to take tests such as GCSEs when they are ready, and to progress through different academic subjects at different rates. In the USA, the Advanced Placement system allows students to take university-level courses in school, be tested, and to have success rewarded by college credits – so a student might enter the second year university course, for example. The Tomlinson Report (2004) argues for a more differentiated curriculum.

Adaptive testing: in some circumstances, the group to be tested is heterogeneous as in the case of language testing, and selection tests for employment. Systems of assessment that change the tasks taken in the light of progress so far can be useful in such circumstances. The principle is straightforward: candidates are presented with tasks of intermediate difficulty; if they are successful, the difficulty level increases; if they are unsuccessful, it decreases. This allows a more accurate estimate of the level of attainment. Adaptive tests can work well when there is a single scale of difficulty – for example in number skill, or vocabulary. They require careful development when a number of different factors affect performance (such as technical as well as problem-solving skills), and are unlikely to be useful where extended responses are required, because the adaptive system has too little to work on. Examples in the school system can be found in Victoria, Australia (AIM Online 2003), where adaptive tests of English and mathematics are used.

Better immediate feedback: candidates can often be given information immediately about success, as is the case in the tests that all trainee teachers are required to take in English, mathematics and ICT (Teacher Training Agency 2003). (This is not necessarily an advantage, if this testing method encourages an 'instrumental' approach, where students learn in order to pass tests rather than to learn things. It could also force assessment design to focus on objective knowledge rather than the development of process skills, if immediate feedback became a requirement for all testing.) In principle, candidates could also be given diagnostic information about those aspects of performance most in need of improvement.

Motivational gains: there are claims (Richardson, Baird, Ridgway, Ripley, Shorrocks-Taylor and Swan 2002; Ripley 2004) that students prefer e-assessment to paper-based assessment, because the users feel more in control; interfaces are judged to be friendly; and because some tests use games and simulations, which resemble both learning environments and recreational activities.

Better exemplification for students and teachers: posting examples of work which meets certain standards can be beneficial. In South Australia, excellent student work in technology is displayed on the web (see www.ssabsa.sa.edu.au/tech/2004techsho/index.htm).

Better 'system' feedback: having full sets of response data from students available at the time of Examiners' Reports can improve the quality of feedback. Details of questions, and parts of questions, that proved relatively difficult and easy should improve the quality of Examiners' Reports (which are based currently on examiners' experiences of a sample of scripts, and rarely on candidate success on questions and part-questions). This information will be useful for both improving the quality of questions, and in providing information to teachers about topics that have not been learned well.

Faster information for higher education: universities need assessment results in a timely fashion. UK universities receive A-level results quite late in the academic year, and engage in a frenetic process to fill places with appropriately qualified applicants when students do and do not achieve the grades that were a condition of entry. These pressures would be eased if results were delivered earlier.

Better task design: it is easier for test constructors to change tasks on the basis of information during testing and pre-testing, because of the immediacy of data collection. This can range from the rejection of items that do not function well (for example items where students who score well overall are likely to fail a particular item) to improved test design (for example, ensuring that there are a lot of items set around critical cut-off points – especially the pass/fail boundary – so that the test is most reliable there).

Cost: it is common to claim that e-assessment can save money – it is clear that online multiple choice tests can be cheap to administer and score. However, if we are to exploit the potential of ICT to improve assessment – for example by presenting simulations or video as an integral part of a test – then the costs of testing are likely to increase.

3.2 USES OF E-ASSESSMENT TO SUPPORT CURRENT EDUCATIONAL GOALS

3.2.1 Using ICT to support Multiple Choice Tests

This is a well-established technology, particularly well suited to assessing declarative knowledge ('knowing that') in well-defined domains. Developing tasks to identify student misconceptions is also possible. It is harder to assess procedural knowledge ('knowing how'). MCT is unsuited to eliciting student explanations, or other open responses. MCT have the great advantage that they can be very cheap to create and use. Some of this cheapness is illusory, because the costs

CURRENT DEVELOPMENTS IN E-ASSESSMENT

of designing good items can be high. Over-use of MCT can be very expensive, if it leads to a distortion of the curriculum in favour of atomised declarative knowledge, divorced from conceptual structures that students can use to work on the world, effectively. MCT are used extensively in the USA for high-stakes assessment, and are presented increasingly via the web. For example, web-based high-stakes State tests are available in Dakota and Georgia; the Graduate Record Examination (GRE), used by many colleges to determine access to Graduate School in many US colleges, is available online.

it makes sense to allow students access to the tools they use in class, during testing

3.2.2 Creating more authentic paper and pencil tests

It makes sense to allow students access to the tools they use in class, such as word processors, and that professionals use at work, such as graphing tools and modelling packages, during testing. It makes no sense at all to always forbid students to use 'tools of the trade' when being assessed. E-learning changes the nature of the skills required. E-assessment allows examiners to focus more on conceptual understanding of what needs to be done to solve problems, and less on telling students what to do, then assessing them on their competence in using the manual techniques required to get the answer. In Australia, the State of Victoria (www.vcaa.vic.edu.au/prep10) has a system for essay marking where students key in their responses to questions, which are then distributed electronically and marked by human markers. Computer Algebra Systems (CAS) can be used in the Baccalauréat Général Mathématiques examination in France; the International Baccalaureate Organisation (IBO) is running a CAS pilot for its Higher Level Mathematics Diploma from September 2004. In the USA, CAS can be used when taking the College Board's Advanced Placement Calculus test.

3.2.3 Using ICT to support current UK examination processes

A number of ways in which ICT can improve current examination practices are set out below.

Better school-examination board communication: Tomlinson (2002) points to existing extensive use of ICT by awarding bodies in the examination process, and argues for more use of Electronic Data Interchange (EDI) systems, which enable schools and colleges to submit examination entries and information about candidates online and to receive results automatically.

Supporting the current marking and moderation process: a challenge faced by large-scale tests that require human markers is to ensure the comparability of standards across markers, and over time for all markers during the grading process. Chief examiners create scoring rubrics to guide other markers, and there is usually a process of standardisation where markers use the scoring rubrics to score a sample of scripts, and attend a standardising meeting where standards are compared, discrepancies are discussed, and the rubric is tuned. Once markers have reached an appropriate level of marking accuracy, they mark examinations independently. Systems vary in terms of the extent of the moderation used. In some systems, scripts are sampled by chief examiners, and serious deviation from the

rubric can lead to the remarking of all the scripts sent to a particular examiner. ICT can be used to support this process. Sample scripts typical of different categories of student work can be put online, for easy reference by markers. Entry of marks can be done via templates that ensure that markers complete every section, and the tedious process of aggregating marks from different parts of the script is done automatically and without error. Data is collected in a way that facilitates rapid and detailed analysis, at the level of responses to different parts of questions, whole questions, and the distribution of test scores.

Replacing paper: in the USA (and increasingly in the UK), there is widespread use of systems where students take paper-based examinations, and the scripts are scanned electronically (this is analogous to Optical Mark Recognition for multiple choice tests that has been available for many years). Once in this format, the documents can be sent electronically to markers, who can be working almost anywhere. These systems have a number of advantages over paper-based systems. First, there are considerable problems in tracking the distribution and return of large volumes of paper to and from markers; there are security issues sending examination papers by post, and scripts can get lost. Second, moderation of the quality of scoring can be done easily. Pre-scored 'anchor' papers can be sent to markers during the course of their marking, to ensure they are maintaining standards; markers who do not perform adequately can be told to take a break, or can be removed from the pool of markers. The whole process can be monitored in terms of the rate at which scripts are being marked. There is flexibility in the ways that scoring is done. Markers can be asked to score whole scripts, or individual questions. So a newly appointed marker might be sent questions judged to be easy to mark, and more experienced markers might be sent questions which require deeper subject knowledge. The reliability of scoring can be increased. Scripts judged to be around key borderlines on first marking can be sent to other markers; scripts judged to be well away from boundaries need be scored only once. Online support can be provided; markers can ask for help with specific student responses. Data is captured in a form suitable for a number of subsequent analyses.

An interesting variant of this approach that obviates the need for scanning would be to require candidates to use 'intelligent pens'. These pens have two distinct functions. The first is to write like a conventional pen. The second is to record its movements (exactly) on the page. This is done by using specially prepared stationery. Imagine you could see a small square area of a banknote. The pattern across the whole surface is never repeated, so that, given sufficient time, you could find exactly where the square is located on the note. The pen works in a similar way, to record its position on the page over the course of the examination. The pen is then connected to a computer, and all the data is downloaded. The whole student response can then be reconstructed. Clearly, this approach would have to be subjected to extensive trialling before any widespread adoption.

CURRENT DEVELOPMENTS IN E-ASSESSMENT

ICT can be used to moderate human markers

3.2.4 Online assessment: turning a GCSE paper into 'computer-only' e-assessment

An interesting challenge is to devise ways to replace paper-based tests with ICT-based tests, and to score them automatically. Some virtues of paper-based tests are unlikely to be replicated for a number of reasons, so setting tests on-screen is likely to bring about changes in the nature of what is assessed. Here, we consider one specimen GCSE mathematics paper to illustrate the problems.

Measuring and drawing: about 10% of the marks in the paper-based assessment required the use of actual 'instruments' (ruler, protractor, compasses). One approach for translation onto screen would be to simulate the physical instruments, eg to provide a virtual protractor that can be dragged around the screen and rotated. Another is to provide CAD or interactive geometry packages. The latter would require a substantial change to the syllabus, but could provide real benefits in terms of student learning.

Mathematical expressions: about 20% of the marks required the student to write down answers that could not be keyed in, using a standard keyboard. These included fractions, division expressions, and powers.

Rough work and partial credit: almost every question in the paper format included space for rough work, and about 30% of the total marks potentially could be awarded based on this work, in the form of partial credit awarded where the final answer is incorrect (these marks are usually awarded in full if the final answer is correct). There are two distinct problems in translating this to a digital format – first capturing the rough work, and second, allocating partial credit. Computer capture is very difficult, given current interfaces; the rules for allocating partial credit would have to be specified in very fine detail for them to be used as part of an automatic scoring routine.

3.2.5 Scoring of open responses

GCSE questions often require students to answer questions in their own way, and to explain things – scoring these responses automatically is inherently difficult. Automated scoring of open student responses is the focus of a good deal of ongoing work. A number of approaches have been taken to the problem of automatic scoring. One is based on the analysis of the surface features of the response (Cohen, Ben-Simon and Hovav 2003), such as the number of characters entered, the number of sentences, sentence length, the number of low-frequency words used, and the like. The success of such methods can be judged by comparing the correlation between computer and human judges, and the correlation between scores given by two sets of human judges. Cohen, Ben-Simon and Hovav (2003) looked at the scoring of a range of essay types by humans and computer, and report that the correlation between the number of characters keyed by the student, and the scores given by human judges are as high as the correlation between scores given by human judges. Nevertheless, these scoring systems do not provide a panacea. In the USA, double marking is used to ensure reliability (this is rarely done in the UK). ICT can be used to moderate human markers (and save money) – if the computer and the human disagree, the

paper is re-marked by a human. Machine-only scoring is unlikely to be useful in UK contexts, for two reasons. First is that the UK culture requires that scoring schemes be described in ways that are useful to teachers and students. Second is that the consequential validity of such scoring systems would be dire – the advice to students would be to improve their scores simply by using more keystrokes. A second approach which could improve the quality of scoring and reduce costs is being used to assess student responses on tasks in contexts where the range of acceptable responses can be well defined, such as in short answer science tasks (eg Sukkarieh, Pulman and Raikes 2003). Here, appropriate ('the Earth rotates around the sun') and inappropriate ('the sun rotates around the Earth') responses are defined. Lists of synonyms are generated for nouns ('our globe') and verbs ('circles'), and alternative grammatical forms are defined, based on analyses of large numbers of student responses. Student responses are parsed using techniques borrowed from Natural Language Processing, and are compared with stored appropriate and inappropriate responses, using a variety of Information Extraction techniques (see Cowie and Lehnert 1996). Mitchell, Aldridge, Williamson and Broomhead (2003) describe work at The Dundee Medical School. Here, all students take the same examination at the end of every year. Academics are presented with all the responses to the same question, with the computer's judgement on the correctness or otherwise of the answer, and an estimate of the confidence of the judgement. Human scoring time is dramatically reduced, and staff report positive benefits in terms of the quality of the questions they ask, both in terms of rewriting ambiguous questions (which produce student responses that are difficult to score) and in terms of writing questions which highlight student misconceptions. This approach requires a good deal of work prior to live testing, so is well suited to situations where tasks will be used repeatedly.

In the USA, the Graduate Management Aptitude Test (GMAT) - used to determine access to business schools - uses automated scoring of text. Here again, the test is scored by both human and machine, to offer some sort of reliability check for the human marker.

3.3 ICT SUPPORT FOR CURRENT 'NEW' EDUCATIONAL GOALS

There is an emerging consensus worldwide on 'new' educational goals, focused on problem solving using mathematics and science, supported by an increased use of information technology (compare, for example, UK developments with those in New Zealand www.minedu.govt.nz; and Singapore www1.moe.edu.sg/iteducation). These new goals involve the development of higher-order thinking, and a range of social skills such as communication, and working in groups. There is an honourable tradition of assessing problem solving via the use of extended tasks, such as those developed by the APU (eg Archenhold, Bell, Donnelly, Johnson and Welford 1988). However, the computer offers some unique features in terms of representation, interaction, and its support for modelling. Here, we describe some recent developments which make use of these unique features.

new goals involve the development of higher-order thinking, and a range of social skills

CURRENT DEVELOPMENTS IN E-ASSESSMENT

3.3.1 The development of World Class Tests

Tests were designed to identify high-attaining students in problem solving in mathematics, science and technology at ages 9 and 13 years, as part of the work on the World Class Arena (www.worldclassarena.org). Computers make it easy to present new sorts of tasks, for example tasks where dynamic displays show changes in several variables over time, or which present video of a situation which students must model. A wide variety of representations can be supported, and students can be asked to switch between them. The interactive properties of computers make them well suited to the assessment of process skills.

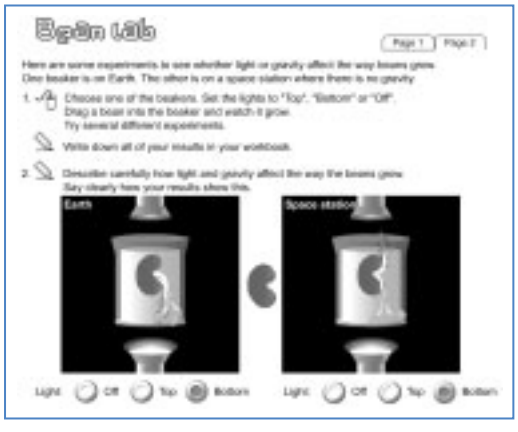
Using computers to give students control over how data is presented allows them to work with complex data sets of a sort that would be very difficult to work with on paper. Tasks can be set in realistic contexts, using realistic data to address problems of considerable complexity, using resources and methods that are familiar to professionals working in the relevant field. Two examples are presented here: Oxygen and Bean Lab.

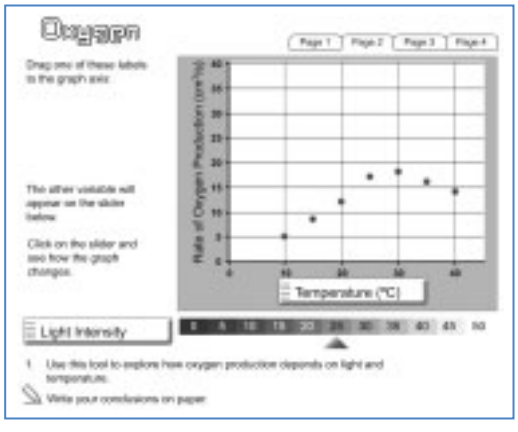
the interactive properties of computers make them well suited to the assessment of process skills

Further examples of tasks can be found in Ridgway and McCusker (2003). Skills assessed include:

Understanding and representing problems: traditional educational goals such as the ability to interpret tables and graphs, and to translate information coded in one representation into information coded in another representation continue to be vital skills for mathematical and scientific literacy. Computers allow fast and reversible transformations of information from one representation to another, and students can be asked to explain the relationships between them.

Assessing process skills in science and mathematics: the desire to assess process skills is not new. Traditionally, students would be presented with tasks in laboratories, or would be required to keep logs and portfolios of their laboratory work. However, the laboratory setting can introduce elements which reduce the reliability of the assessment, such as instruments which fail to function properly, or materials whose properties are less than ideal. Students are required to physically manipulate apparatus – chance differences between students in terms of





their previous exposure to particular equipment can both reduce reliability, and add an extra cognitive load to the intellectual task being performed. In some situations, issues of health and safety arise. Some education systems are unwilling to accept teacher ratings of students for the purposes of high-stakes testing, with the result that process skills in science are not assessed at all. Computer-based assessment permits the assessment of these valuable aspects of learning science, at modest cost. A range of different process skills can be identified, which include:

- working systematically (for example, choosing tests systematically, controlling variables and recording results systematically)
- generating and testing hypotheses
- finding rules and relationships
- handling complex data
- testing solutions
- seeking completeness and rigour (in many real-world situations, exemplified by diagnosis and remediation in spheres such as medicine and industrial process control, it is important to find all of the faults in a system).

Five sets of live tests have been administered in the UK and elsewhere, each of which was preceded by extensive pre-testing. A notable result was the ease with which students interacted with computers. The affective response from students was very strong – they really enjoy working on these tasks. This might be related to the sustained challenge the tasks present, which is similar to the reported reasons why they like computer-based games (Kirriemuir and McFarlane 2004).

Students performed better on some tasks than one might expect – notably tasks that require them to reason from complex data sets (eg data with two independent variables and one dependent variable at age 9 years). We take this as a very positive sign that computers can play a leading role in the development of the skills which constitute the new educational agenda. In many aspects, student performance was poor - work characterised by guessing, too little use of systematic methods, poor hypothesis generation, and poor generalisation. On many tasks, students were able to show evidence of good reasoning skills; however, explanations were often weak. Given the earlier discussion of the impact of assessment on the curriculum, it is to be hoped that the use of e-assessment of process skills will lead to better student performance on a range of important activities.

World Class Tests focused on summative assessment in science, mathematics and technology, and used a variety of contexts, including geography and economics, as well as biology, physics, and engineering. The ideas are generic, and can be applied to many curriculum areas. On the basis of analyses of student performance on WCT, teaching modules for whole class use have been developed, targeted on weak process skills. These teaching modules provide a good deal of formative assessment, and require students to engage in reflective activities such as critiquing student work, and explaining their own solution strategies.

We discuss 'new' educational goals that are less amenable to summative assessment – such as the ability to work in groups, to communicate, to learn to learn – in Section 4.

computers can play a leading role in the development of the skills which constitute the new educational agenda

CURRENT DEVELOPMENTS IN E-ASSESSMENT

assessment systems must require students to show the full spectrum of competencies

3.3.2 Assessing ICT at Key Stage 3

Ongoing work funded by QCA sets out to assess student attainment in ICT at age 13 years. A key principle for the design of these tests is that students should be tested on their performance on extended tasks ('create a web page about topic X for audience Y, using a particular set of resources - a database, 'clients' accessible via e-mail, spreadsheets for planning, web page creation tools') not on a series of sub-tasks ('use a spreadsheet to add up these numbers'). An extraordinarily ambitious goal is to present tasks and score performance entirely by computer. This is a laudable aim, and shows a government commitment to high quality e-assessment (including £20m for the project).

3.3.3 Digital portfolios

An historical legacy which bedevils the current education system in the UK is the distinction between 'academic' and practical' subjects. This was enshrined in the 1944 Education Act, which created grammar, technical and secondary modern schools (Tattersall 2003). Abstract thinking is important; appropriate action in context that rests on practical competence is important. Neither is much use on its own, and students should be taught to both abstract and apply. For this to become a classroom reality, assessment systems must require students to show the full spectrum of competencies in a number of school subjects. If high-stakes assessment systems fail to reward such behaviours, they are unlikely to be the focus of much work in school. E-portfolios offer a way forward.

There are three distinct uses for portfolios. The first is to provide a repository for student work; the second is to provide a stimulus for reflective activity – which might involve reflection by the student, and critical and creative input from peers and tutors; the third is as showcase, which might be selected by the student to represent their 'best work' (as in an artist's portfolio) or to show that the student has satisfied some externally defined criteria, as in some teacher accreditation systems (eg Schulman 1998). These uses are not mutually exclusive. Students may well wish to archive all their work; reflective activities and feedback from others will be based on a subset of this work; the final 'presentation portfolio' will be selected from this corpus.

These different uses of portfolios reflect different, but not always incompatible, theories of learning. A behaviourist approach will focus on defining 'core competencies' that are impossible to assess in timed examinations, and the need for fast and efficient feedback on student products. A social constructivist view will focus on the importance of reflection and sense making by a group (including the tutor) which will include the negotiation of educational goals.

ICT provides an opportunity to introduce manageable, high quality coursework as part of the summative assessment process. Student portfolios have been advocated for a long time, and have been used on a limited basis. From the viewpoint of assessment, the rationale for portfolios is clear: there are a number of valuable activities and attainments that cannot be assessed using the format of timed tests. The ability to create, design, reflect, modify and persevere are all

important goals of education. It is entirely appropriate to assess these processes by collecting evidence on the ability to engage in an extended piece of work, and to bring it to a successful conclusion by the creation of some product – lab report, video, installation etc. Part of the portfolio can (should) provide evidence of the range of personal skills demonstrated, perhaps under the headings suggested in the Tomlinson Report (2004): student self-awareness – of themselves and the ways they learn and what they know; how students appear to, and interact with, others; thinking about possible futures and making informed decisions. A section of the portfolio in the form of a viva, or simply annotations of products where students show their attainments in these three aspects of performance is appropriate.

A number of problems are associated with portfolios and other sorts of coursework. One is the problem of storage – especially in design projects and in art. ICT can solve the problem by holding images of artefacts created. A second problem is student misbehaviour; this can have a number of forms. One is simply that work is plagiarised; another is that students create some artefact, then 'back-fill' by inventing the development process (which is often assessed as part of the final mark) post hoc. ICT can help with both of these problems by requiring the submission of images of intermediate products, with time stamps. On a more positive note, the ability to store and work with images (photographs, video) is likely to make teaching of the design process more effective. Devices such as mobile phones with in-built cameras and facilities for audio recording make it easy to document the evolution of ideas and artefacts. This facility serves a number of functions. First, it simplifies the documentation of the development of work – reducing the 'busy work' students might otherwise have had to engage in. The process of documentation via a portfolio of work supports student reflections on processes – on decisions made deliberately, those forced by circumstances, and those that just sort of happened. Digital images are easy to manipulate and present. Student presentations of work on the development of artefacts is easy, once images are captured digitally.

In some subjects, such as design and technology, and art, extended projects are at the heart of the discipline. The use of e-portfolios maps directly onto current conceptions of the domain, and offers practical solutions to some common problems (eg Kimbell 2003). This work is important, and is likely to be applicable on a large scale in the near future. A very large number of institutions have made use of portfolio systems; the American Association for Higher Education (AAHE) Portfolio Clearinghouse (www.aahe.org/teaching/portfolio_db.htm) provides an online searchable database of profiles of electronic portfolio projects and resources in higher education, and is a valuable source of ideas.

3.4 SUMMARY OF SECTION 3

There are a number of exciting developments in the use of e-assessment for both summative and formative purposes, and several UK developments are at the leading edge, worldwide. In the UK, the government has decided that extensive use will be made of e-assessment. Some of these developments are a response to current problems

the ability to create, design, reflect, modify and persevere are all important goals of education

CURRENT DEVELOPMENTS IN E-ASSESSMENT

associated with increases in the volume of assessment; some reflect a desire to improve the technical quality of assessment (such as increased scoring reliability), and to make the assessment process more convenient and more useful to users (by the introduction of on-demand testing, and fast reporting of results, for example). E-assessment also makes it possible to assess aspects of performance that have been seen as desirable for a long time – such as the assessment of process skills, and the efficient handling of student portfolios. Using E-assessment to test student ICT capability represents an extremely ambitious goal of presenting holistic tasks to assess performance, rather than a collection of short tasks which are symptoms, rather than exemplars, of ICT capability. Nevertheless, some major challenges face these new developments. Paper tests have a number of advantages in terms of the quality of the image presented, and the variety of ways in which students can respond; automatic scoring of responses will be very difficult, and in some cases impossible to achieve via computer.

A complete reliance on paper-based assessment has a number of drawbacks; first is that such assessments are increasingly 'inauthentic' as classroom and professional practices embrace ICT. Second is that such assessments constrain progress, and have a negative effect on students who have to learn (just for the exam) how to do things on paper that are done far more effectively with ICT. A third major constraint is that current innovative suggestions for curriculum reform, which rely on student portfolios for their implementation, will be impossible to manage on a large scale without extensive use of ICT.

E-assessment is a stimulus for rethinking the whole curriculum, as well as all current assessment systems. E-assessment provides a cost-effective way to integrate high quality portfolio assessment with externally set and marked tests, in any combination. This makes it likely that there will be significant changes in the structure of summative assessments, because of the range of student attainments that can now be assessed reliably. There is likely to be extensive use of teacher assessment of those aspects of performance best judged by humans (including extended pieces of work assembled into portfolios), and more extensive use made of on-demand tests of those aspects of performance which can be done easily by computer, or which are done best by computer.

OPPORTUNITIES AND CHALLENGES FOR E-ASSESSMENT

4 OPPORTUNITIES AND CHALLENGES FOR E-ASSESSMENT

Here, we consider some issues which need to be addressed as a matter of urgency. First are some speculations on how we might assess process skills - essential but often ill-defined educational goals. It will be important to establish the value of such assessments as part of large-scale summative assessment, in contrast to their roles as potentially useful components of formative assessment. It will also be important to establish the appropriate scale of such assessments, and their locus in the curriculum, in terms of educational gains and manageability. Second, we consider the problems of 'going to scale'. Large scale innovation – especially where computers are involved – does not always run smoothly.

4.1 ASSESSING PROCESS SKILLS

4.1.1 Assessing metacognition

As we move towards a knowledge-based society, the development of metacognitive skills increases in importance, and they become educational goals in themselves. Currently, these goals are ill-defined in that there is not yet a consensus in the educational community about their exact nature or how they can be assessed. Goals can be described, and recognised when they are achieved, but exemplification needs further work, and a general sharing of ideas. Ridgway, Swan and Burkhardt (2001) exemplify this process as part of 'Assessing Mathematical Thinking' in materials developed for the US National Institute for Science Education (www.wcer.wisc.edu/nise/cl1).

Here, examples of metacognition are given under four headings: knowing how to use knowledge; analysing and improving cognitive processes; supporting reflection and critical skills; and assessing competence with different thinking styles.

Knowing how to use knowledge: the web offers great opportunities and pitfalls for assessment. Most obviously, the existence of the web means that successful use of it should be an educational target. Expertise in navigation, such as learning how to bookmark useful sources, and how to refine searches are useful skills, but are subsidiary to a set of meta-knowledge skills about the nature of knowledge – how it is constructed, presented, and used by different people for different purposes. There is a need for students to develop sophisticated theories-in-action about knowledge. These theories should include accounts of the nature of knowledge – its generation, and the various functions it serves (including its use as just another rhetorical device!). Students also need to know about their own knowing – what they do and do not know, how they acquire, lose and change their own knowledge – and how they control their cognitive processes when solving problems.

We address the first goal elsewhere in the discussion on assessing competence in ICT. The latter goal is illustrated by Lord Armstrong's remark "power is knowing how to use knowledge". The common corruption to "knowledge is power" misses Armstrong's point almost entirely. Our educational ambitions should be to encourage students to become sophisticated users and creators of knowledge. Good formative assessment should contribute to students' development; web-based sources can

there is a need for students to develop sophisticated theories-in-action about knowledge

OPPORTUNITIES AND CHALLENGES FOR E-ASSESSMENT

be part of both formative and summative assessment of these key elements of student performance.

Key aspects of performance relate to the exploration of the origins of the source, analysis of its qualities as a source, and its relation to a wider set of information. Successful formative assessment helps students to internalise questions and question styles. For summative assessment, we expect students to ask questions about the nature of the information source. The originator can be important – dietary advice from Kellogg's should be treated more cautiously than advice from the British Medical Association. Who created it? For what purpose? From what perspective was this written? The poor quality of much of the information on the web can be a virtue, pedagogically, because students see the sense in challenging the authority of any source, and can do so easily by considering alternative sources (eg Downes and Zammit 2000).

Skills in analysing documents in terms of their style and their use of particular rhetorical devices, and in creating documents for different audiences and in different writing genres, are being developed and used in English (and sociology and philosophy at university level). Again, the ubiquitous use of web sources provides both a rationale for the value of these analytic and creative activities, and a rich source of resources for assessment purposes.

The web makes it easy to compare and contrast different interpretations of 'the same' events by different 'news' providers, and by the same provider over time. In terms of assessment, students can be asked to compare and contrast different presentations, and to describe the evolution of a news event over time. This requires analysis of the way that evidence is selected, and the ways that 'events' are reconstructed over time.

A further key aspect of knowledge use is the ability to relate a particular source to a larger body of knowledge. It will always be important for learners to develop rich schemas of knowledge – facts, skills, and procedures and their interconnections – as the basis for judging the value or otherwise of putative new information, or a theoretical account. In science, a simple example is a digital image of a mammal with horns and claws. Students are expected to say it is most unlikely, because horns are associated with herbivores, and claws with carnivores. At a higher level of abstraction, students might be asked to resolve famous conflicts in scientific ideas, in terms of what was known at the time. For example, Lord Kelvin – probably the most distinguished scientist of his day – argued against the theory of evolution, on the grounds that the timescale was impossible. The core of the Earth is largely molten, but if the Earth were really the millions of years old needed for evolutionary processes to work, it would have cooled down long ago. What didn't he know (or is his criticism valid)? The web is a source of information that challenges current knowledge – students can be asked to relate 'breaking' research to a wider set of knowledge. The recent scare over the MMR vaccine (and the damage that will be done to children by an under-analysed and over-publicised piece of research) provides an example.

A vivid example of summative evaluation which requires both a deep knowledge

schema and powerful skills in knowledge deconstruction and reconstruction is provided by a final undergraduate examination at Goldsmith's University on the art history course, where students are presented with two pictures, side by side, which they are to compare and contrast. They are required to name the artist, deconstruct the iconography, and interpret each work in its historical context. This could be presented via ICT, and could be extended to film, and to other contexts.

Another approach to supporting reflection about knowledge acquisition and creation is to incorporate assignments that require a reflective account of the process of creating some artefact (object or written). Students can be asked process questions about sources of information – ways to find good sources (perhaps in the form of 'advice to someone with a similar job to do'), and about the sources themselves. They can be asked about problems faced, and the ways they were solved, in these 'meta-learning' essays.

'Open-web' examinations offer a parallel to open-book examinations. One virtue of such examinations is that they are more 'authentic' than conventional examinations, in that, outside educational contexts, one rarely has to answer a substantive question without any resources. They allow the examiner to set a broader range of questions, because students are not expected to retain all the relevant information in memory. An adaptive strategy for success on such examinations is to develop meta-knowledge of the whole area, and to index sources very carefully. A large information bank with no index is of little use. Compare the preparation necessary for this sort of examination with the 'cramming' strategy' that can be effective when preparing for conventional examinations. There, the danger is that students hold information in a relatively temporary state for the purpose of the examination, then forget the information once the examination is over. Open-web examinations are likely to have desirable 'consequential validity' – that is to say, are likely to lead to desirable learning (and learning strategies). The unpopularity of open-book examinations (which probably arises because they require serious thought about the subject matter) is likely to apply equally to open-web examinations. The potential for fraudulent behaviour by students (such as e-mailing for advice in situations where the purpose of testing is to assess the ability to search the web, or searching the web when the purpose of the assessment is to assess 'networking' skills) means that student activities will need to be constrained in appropriate ways. Nevertheless, open-web assessment should be explored further.

Analysing and improving cognitive processes: interactive whiteboards can provide the facility to work as a whole class on a problem or simulation, then to replay and critique the sequence of actions. This provides the opportunity to discuss seemingly abstract concepts such as 'strategy' and exemplify them with concrete examples. Analogies with the analysis of games (eg tennis) can make the activity seem natural in class (of course, analysis of on-screen video of ongoing games is a specific example of the sorts of analyses being described here). The long-term intention is to help students develop metacognitive skills that will be applicable in a wide variety of situations. By looking at different solution attempts, students can be asked high-level questions such as

open-web examinations are likely to lead to desirable learning (and learning strategies)

OPPORTUNITIES AND CHALLENGES FOR E-ASSESSMENT

'how do you solve problems of this sort?' – which can be assessed more formally by tasks such as 'write some guidance for someone else, that will help them to solve problems like this one'. A requirement for summative e-portfolios could be that sample reflective analyses of processes be included.

These techniques have great potential when the focus is on the social and emotional education of students. Topics raised in personal and social education such as approaches to bullying can be approached by presenting students with video vignettes, and asking them to describe situations, the interactions that take place, and the feelings of participants. Parallel information channels (provided by the participants) can provide students with feedback on the correctness or otherwise of their insights. At a lower level, assessing children's ability to identify the emotions being expressed in different faces can give insights into their developmental state (or, in more extreme cases, into pathological states such as autism). If summative information is appropriate, it can be based on the analysis of such vignettes.

Supporting reflection and critical skills: an important higher-order skill is the ability to review and improve work. This can be done via paper and pencil (for example by writing on every third line, and changing pen colour at every revision cycle), but is made very easy by the use of ICT, with facilities such as 'track changes' in MS-Word. Students can be asked to provide examples of their ability to improve work on the basis of others' and their own suggestions, and of their ability to critique the work of others. Another way to assess critical thinking is to require students to annotate work to show where they meet the assessment criteria.

Courtenay (personal communication, 2004) described an activity designed to support creative writing in English in a night class comprised of 30 non-native speakers at an early stage of learning English. Courtenay focuses on creation and critique, and seeks to spend as much time as possible interacting with his students. Each student writes online, and when they are satisfied with their composition, it is posted to a shared server. Every student is required to offer constructive comments on five compositions, and to revise their own writing in the light of five sets of comments. The teacher is able to tour and coach individuals as they write. With little effort, this approach could be extended to providing summative assessment. Students could be required to submit their comments on others' writing to be evaluated, and could provide evidence of their ability to use comments on their own work. An assessment system like this would reinforce rather than distort the educational ambitions of the teacher.

Peer assessment is attractive for a number of reasons. (Topping's 1998 review demonstrated that it is associated with gains on conventional performance measures, in higher education.) Students can be asked to create far more pieces of work than could be marked by a single tutor. It can avoid the problem that as a class size gets bigger, the load on the tutor increases directly, along with the time taken to provide feedback to students. Students must understand criteria for assessment, and must acquire a range of higher-order skills, such as abstracting ideas, detecting errors and misconceptions, critiquing and suggesting improvements, if

they are to engage in peer assessment. Peer assessment is a fact of life outside education, so peer assessment is far more 'authentic' than some forms of assessment such as multiple choice tests. Possible disadvantages relate to the possibility of an enhanced workload on students, unreliable feedback, and biased feedback.

A number of commercially available systems have been designed to support peer assessment. Calibrated Peer Review™ (Chapman and Fiore 2001) was designed to support the peer assessment of essays in molecular science, but has been applied in a variety of subjects, and with students across the education system. Students write short essays, and are asked questions designed to foster their critical thinking. Students are presented with three 'calibration' essays to grade, and must demonstrate their competence before they progress. Two of the essays contain errors and misconceptions which students must identify and correct. Students are also asked questions on style and grammar. The scores they give to the assignments are compared with 'official' scores, and a calibration report is created for the student and the tutor. If performance is inadequate, more instruction is provided, and the student must repeat the activity. Once they have shown that they can assess essays effectively and reliably, they are asked to grade three essays by peers, and finally are asked to grade their own essay. The student and the instructor receive comments and scores.

CPR is not restricted to essays in science; the idea is generic, and can be applied to literary criticism, commentaries on a piece of art, or laboratory reports, for example. The tutor must select the focus of the assignment, write an exemplar answer for calibration, and select two pieces of student work which contain interesting errors or omissions. Each of these has to be graded by the tutor, and relevant comments have to be written. The tutor also writes key questions on content and style. CPR is designed to overcome the potential weakness of peer assessment in terms of unreliable assessment (via training and moderation) and bias (via anonymity). The authors claim considerable gains in students' ability to 'learn to learn' because their attention is focused on abstracting ideas and arguments, describing, analysing and assessing the quality of material, and in review. CPR also increases the amount of writing that students do.

Doiron and Isaac (2002) have developed a novel form of online peer review designed to complement the American College of Surgeons Advanced Trauma Life Support Course for fourth year medical students. Their system involves self-assessment, peer evaluation, feedback and debate. There is an inherent problem giving large numbers of students direct experience of Emergency Room procedures. Here, students are presented with a realistic case study, and must prevent the patient from dying, conduct clinical tests, then request appropriate lab work followed by diagnosis and recommendation of a treatment. Students reflect on, and self-assess, their knowledge. They submit a diagnosis and proposed treatment plan to the whole group. For peer review, they are presented with two other diagnoses and treatments – one from the tutor, prepared to contain errors, for critique. If the student fails to detect the errors, they get individual feedback from the tutor. Students then review 'live' reports from

peer assessment is a fact of life outside education

OPPORTUNITIES AND CHALLENGES FOR E-ASSESSMENT

mobile phone technology might provide a means of assessing thinking styles

two of their peers (so three reviews are considered together). Where there are disagreements, the two views are presented to a larger group (four to ten students) who must all offer their own view, and debate the issue. Similar work is being conducted on a health psychology course, and in engineering.

Assessing competence with different thinking styles: mobile phone technology might provide a means of assessing thinking styles via simulated group work. Here, each student works in a simulated environment, where responses from other 'group members' are pre-specified, and some responses to the actions of the student are pre-defined. This environment is artificial for a number of obvious reasons – contact is via phone (or e-mail) rather than face-to-face and the range of dynamic interactions is constrained. However, these constraints mean that students can be assessed in relatively standardised conditions, and sequences can be replayed for analysis and reflection as part of formative assessment.

Analysing the ability to engage in De Bono's (2000) 'Thinking Hats' activity provides a concrete example. De Bono has identified a number of thinking styles, all of which are useful when solving problems. None is effective on its own. He argues that people differ in their preferences for these different thinking styles, and often stick with a particular style of thinking. In terms of group dynamics, individuals can become ego-involved with a particular style of thinking, with negative consequences for the productivity of the group. De Bono argues that these different thinking styles should be made explicit, and that every group member should engage with every thinking style in the course of group work. He suggests a formal mechanism for this, where thinking styles are associated with hats of different colours, and group members are invited to take particular roles – sometimes as individuals, and sometimes as a whole group. Thinking styles include asking about what is known or what is needed (the White Hat); saying why an idea won't work (the Black Hat); generating ideas and alternatives (the Green Hat); describing feelings, hunches and intuitions (the Red Hat); managing group processes (the Blue Hat); and the optimistic advocacy of ideas (the Yellow Hat).

Given some specific suggestions for actions via mobile phone or e-mail, students can be asked to work in Red, Yellow and Black Hat styles; or given a stream of (simulated) input to a conference, students can be asked to work in Blue Hat mode. Their responses provide information on their strengths and weaknesses working in different thinking styles. This idea is not restricted to de Bono's framework, but is a generic idea for assessing individual skills in group settings.

4.1.2 Assessing group projects

A valuable skill is the ability to work productively in groups. This requires good communication skills, understanding the criteria for effective group work, understanding different roles, the ability to assess one's own work and the work of others, and the ability to respond positively to formative and summative feedback. The assessment of group work is problematic for a number of reasons: problems can be caused by 'social loafing' and the allocation of equal marks for unequal

contributions; undesirable effects of students rating peers; and time-hungry procedures for gathering accurate evidence on student performance.

SPARK (Self and Peer Assessment Resource Kit - www.educ.dab.uts.edu.au/darrall/sparksite) is an academic open source project designed to support the effective evaluation of group work, that has been used in a variety of contexts in higher education. It requires a clear specification of the tasks to be performed by the group and the assessment criteria. Students reflect on group processes during the performance of the task, and rate all the group members, and themselves against the criteria provided. The tutor monitors the work of the group, grades the product of the group work, uses SPARK to convert group marks into individual marks, and provides individual summative and formative feedback (eg that a student rates their own contribution to the group far higher than other group members do). Evaluations of SPARK by its authors in a variety of higher education contexts have been positive (eg Freeman and McKenzie 2002).

4.1.3 Assessing creativity

'Creativity' involves the production of a new idea or artefact that is judged by some community to be of value. Many writers have made a distinction between analytic and creative thinking. Analytic thinking has been characterised as: linear, rational, logical, conscious and deliberate. Creative thinking has been described as: parallel, unconstrained, illogical, unconscious, and chaotic. Creativity became a bandwagon for education in the 1960s, in part as a healthy corrective to an over-emphasis on 'Intelligence'. A problem with some of these early proponents of 'creativity' (eg Getzels and Jackson 1962) was that they accepted many of the philosophical assumptions of the Intelligence movement, and many of their methods, but were incompetent in their use. The result was a movement that was based on some good ideas, but which was poorly theorised, and supported by flawed evidence. Just as there are many styles of analytic thinking, that are coloured and improved by knowledge in particular domains, and different ways to represent information, so too are there many styles of creative thinking, again, influenced by knowledge and experience in a variety of domains. Creativity (as defined above) requires an intimate interplay of creative and analytic thinking. It is important to develop creativity, and to evaluate the products of creative thinking. Creativity should be evaluated by an analysis of product, and by an analysis of student processes, using methods described earlier (notably, tracking the design process, and reflective accounts on this process).

It can be difficult to obtain good paper-based accounts of student processes and results after engaging with an extended piece of work. This can be a desirable activity for a number of reasons. First, it requires students to translate knowledge from one form to another, and to consider the needs of a different audience – notably from a static written form whose primary audience is the teacher, to a visual and dynamic form for some predefined audience, who will have a range of understandings about the topic in hand. Second, it is inherently valuable as a skill. Digital cameras and whiteboards make it easy for students to show their work (which might be on paper, in the form of

it is important to develop creativity, and to evaluate the products of creative thinking

OPPORTUNITIES AND CHALLENGES FOR E-ASSESSMENT

manipulatives, or some artefact that has been created) and to explain what they have done, justify their answer, and describe the design decisions they took.

4.1.4 Assessing communication skills

Mobile phones could be used more extensively for assessment. A simple example would be to use mobile phones for the aural comprehension aspect of language learning. Current practices of using an analogue tape recorder at the front of a classroom are inherently unfair. The quality of the sound will differ as a function of the tape machine used; the sound intensity at the front of the room will be dramatically higher than at the back of the room. Using conventional computer technology, Southern Australia uses MP3 files to test language comprehension (see www.ssabsa.sa.edu.au) – clearly, good practice.

The eVIVA project (www.qca.org.uk/adultlearning/downloads/eviva_project.pdf, www.eviva.tv) uses phones as the medium for oral testing with portfolio-based Key Stage 3 ICT assessment. Students can book a test session, and so can have (almost) on-demand testing. The phones are also used for recording 'voice postcards' of learning milestones, and posting these to a central website. The 'voice postcards' can be used by a student to support the piece of portfolio evidence which they are presenting.

As speech recognition technologies continue to improve, one can envisage a situation where questions are posed orally by telephone, and student responses are scored automatically. In the case of language learning, this could be applied to elementary aspects of learning such as pronunciation, to vocabulary, and to correcting sentence structure 'mistakes' presented to students. Given test technologies that support 'tailored testing', the phone system could be used to provide on-demand testing of some aspects of language use. Such systems are unlikely to be useable (in the short term at least) for high-stakes testing, because of problems of impersonation. These problems may be removed if effective person recognition systems are developed and introduced on a large scale.

4.2 NATIONAL CURRICULA, NATIONAL ASSESSMENT

The Tomlinson Report (2004) addresses fundamental questions about curriculum design and assessment, and describes a number of serious problems with current systems. Assessment exemplifies educational goals, and has a major effect on educational practice. Unless assessment systems are aligned with educational goals, they will distort curriculum ambitions. There is a general desire for more school-based assessment, and more process-based assessment, and an insistence that current high standards of equity and probity in the examination process are maintained. E-assessment (eg via e-portfolios) can provide the means to empower teachers and schools, while ensuring that high standards of assessment are met. ICT can support the whole process of teacher preparation, and the establishment of procedures to ensure comparability of standards across schools. School-based judgements could be moderated by external computer-based tests. E-assessment can extend the range of reliable assessments that can be

conducted, and so can widen the debate on curriculum and assessment design. On-demand testing will have considerable implications for curriculum planning. Students could take summative tests at different times, and could progress through the curriculum at different rates.

E-assessment could reduce the damage caused by current tests. At present, new SAT papers are created each year, and all students answer the same questions. If the purpose of testing is to establish the performance of some system (such as a school or an LEA), better methods could be employed. If there were a large bank of tasks available in electronic form, and different students received a different set of tasks, then coverage of the curriculum could be better, and there would be no need to report individual student scores. This would have the advantage that a larger variety of task types could be used, and would avoid the current distortions caused by teachers 'teaching to the SAT'.

4.3 EVOLUTION AND REVOLUTION

Even where there is a shared vision on future curricula, there can be considerable problems in implementation. Ridgway (1998) draws analogies between ecological restoration and educational change, and describes the sorts of research needed for successful change. This style is close to research in fast-changing fields such as electronics, where discoveries and inventions drive practice and theory, in contrast to well-established fields where theory can lead practice. It is important to be aware that some goals are easy to achieve from most starting points, whilst others need a good deal of capacity building before they can be reached. It will be important to phase the introduction of e-assessment in such a way that the load on students, teachers, schools and systems is lower than the current assessment load. Some barriers are discussed below.

Establishing the credibility of e-assessment: in some areas such as competency-based assessment, the case for e-assessment is self-evident. In other areas, reasonable sceptics will have to be convinced of its value. They will have concerns about the construct validity of new tests (exactly what do they measure?); the reliability of new tests in comparison with existing tests; and the educational standards required – both in relation to current tests, and across tests such as those given 'on-demand' in different places and at different times. Each of these questions will need to be addressed for each family of e-assessments, usually by means of an empirical study.

Building system capacity: there is an urgent need to build capacity for e-assessment that ranges from test design, test delivery and processing, and expertise in school. Each of these is problematic.

Task and test design: very few people have expertise in creating e-assessments, in comparison to the large numbers of people competent to create conventional tests. There is an urgent need to create new task types and to explore their reliability and validity. If we do not continue to explore, students will be faced with a set of tasks which recently were innovative, but which are now hackneyed.

e-assessment could reduce the damage caused by current tests

OPPORTUNITIES AND CHALLENGES FOR E-ASSESSMENT

it is important that e-assessment does not create a 'digital divide'

Establishing technical standards: currently, there are three sets of technical standards. We need a consensus document. The needs of students with special needs must be addressed. Standards for monitoring the quality of the assessments given in schools (actually a rather hostile environment for ICT, because of the plethora of machines and operating systems), and the procedures put in place by examination authorities need to be written, and validated in practical settings.

ICT infrastructure: good broadband systems are needed – in particular, very high specification systems are needed for big schools. Currently, about 40% of primary schools, and about 100% of secondary schools have broadband access, but not necessarily at the levels needed for online assessment (Rt Hon Charles Clarke MP 2004). The proposals set out in the Tomlinson Report are only feasible if a national database of student achievement is established. At school level, extensive investment in ICT will be needed, and costs will recur.

The examination process: dealing with e-assessment poses serious challenges to paper-based examination authorities. They need to develop a robust technology infrastructure, and (at least as important) the competencies of staff to make these systems function effectively. A good start has been made here, for example in the work on the assessment of basic and key skills. However, there are salutary messages from the QCA Report on implementation (QCA 2004). AQA report (Adams and Hudson 2004) that their surveys show considerable satisfaction from examiners. Examiners report that the software is easy to use; they like the increased accuracy and validation at input, and the auto-totalling of marks by the computer, and the electronic management of reporting and discrepancies.

On examiners and examining: High quality training is an essential aspect of reliable assessment. Tomlinson recommends (paras 134–136) "a thorough professionalisation of the role of markers and examiners, including coursework markers", and the Report makes a number of specific recommendations on how this might be institutionalised via schemes for professional development, accreditation, and appropriate professional reward systems. The Secondary Heads Associations have argued for the establishment of 'Chartered Examiners' in schools and colleges, who would give their organisations the right to take more control over examination assessment.

School and test-centre expertise: this presents a massive challenge for professional development. Schools need to develop systems which are robust.

Plagiarism: poses a major threat to all assessment systems (eg Ridgway and Smith 2004). These threats range from downloading work direct from the internet, commissioning work, and impersonation. Assessment systems will need to be resistant to such attacks.

Equity issues: it is important that e-assessment does not create a 'digital divide' which privileges some students over others on the basis of opportunities of access.

4.4 RELIABLE TEACHER ASSESSMENT VIA E-PORTFOLIOS

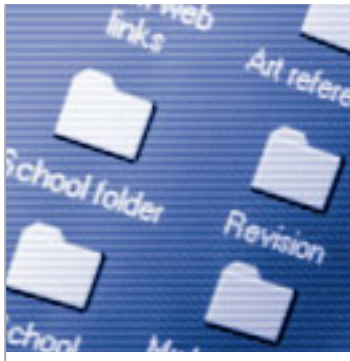
A key decision for educational systems is to decide exactly how much of the students' time should be devoted to working on extended projects, and how much should be based on shorter activities. A related decision is the balance to be struck between portfolio systems assessed in school, and timed external assessments. A key issue is to establish robust and reliable systems of school-based assessment. It is worth highlighting the extreme positions that different systems use. In some systems, all assessment is done externally. In some systems – for example Queensland, Australia - all assessment is school-based. Queensland provides extensive systems for training teachers, and for moderating their judgements. ICT can facilitate this process. All student submissions can be put onto the web, and systems of cross-moderation can be established. Externally defined tests can be used to guide the moderation process.

4.5 DUMBING-DOWN ASSESSMENT

There is a danger that considerations of cost and ease of assessment will lead to the introduction of 'cheap' assessment systems which prove to be very expensive in terms of the damage they do to students' educational experiences. At the time of writing, this seems most unlikely in the UK. QCA have funded some innovative e-assessment developments at investment levels beyond the reach of most companies, and have a large group focused on developing and sharing expertise in e-assessment (www.qca.org.uk).

4.6 SUMMARY OF SECTION 4

New educational goals continue to emerge, and the process of critical reflection on what is important to learn, and how this might be assessed authentically needs to be institutionalised into curriculum planning. In this section, we explore ways to assess metacognition, group projects, creativity and communication skills. E-assessment is certain to play a major role in defining and implementing curriculum change in the UK. There is a strong government commitment to e-assessment, and good initial progress has been made. Major challenges of 'going to scale' have yet to be faced. A good deal of innovative work is needed, coupled with a grounded approach to system-wide implementation.



e-assessment is certain to play a major role in defining and implementing curriculum change in the UK

GLOSSARY

ACKNOWLEDGEMENTS

We wish to thank a number of people who have commented constructively on this document, in particular Keri Facer, Annika Small, Jeremy Tafler, and Kathleen Tattersall. We are grateful to them for their input. All the faults and errors of omission are our own.

GLOSSARY

Adaptive testing a sequential form of individual testing in which successive items in the test are chosen based primarily on the psychometric properties and content of the items, and the participant's response to previous items

A-level (AS/A2) General Certificate of Education (GCE) Advanced Level. Study usually consists of a two-year academic course and students will usually select two or three subjects from subjects studied at AS-levels to continue to A-level (called A2)

Anchor(s) a sample of student work that exemplifies a specific level of performance. Markers use anchors to score student work, usually comparing the student performance to the anchor

AQA an awarding body: Assessment and Qualifications Alliance formed from the merger of Associated Examining Board (AEB) and the Northern Examinations and Assessment Board (NEAB) in 2000

AS-levels General Certificate of Advanced Supplementary Level, considered to be the equivalent of half an A-level. Young people are now expected to study four AS-levels during Year 12 at school or college

Assessment any systematic method of obtaining evidence from tests, examinations, questionnaires, surveys and collateral sources used to draw inferences about characteristics of people, objects or programs for a specific purpose

Basic skills the ability to read, write and speak in English and use mathematics at a level necessary to function and progress at work and society in general

CAS Computer Algebra System. Software package used for the manipulation of mathematical formulae. Automates tedious and sometimes difficult algebraic manipulation tasks. Systems vary and may include facilities for graphing equations or provide a programming language for the user to define their own procedures

City and Guilds major awarding body for vocational qualifications in the UK

Competency-based assessment assessment process based on the collection of evidence on which judgments are made concerning progress towards satisfaction of standard performance criteria

Concept map the arrangement of ideas into a visual layout highlighting connections between associated ideas, revealing the structural pattern in the information

Criterion referenced assessment assessment linked to predefined standards. (eg 'Can swim 25 metres in a swimming pool')

CSE Certificate of Secondary Education: former system of British examinations taken in a range of subjects, usually at the age of 16

Diagnostic testing testing used to identify the conceptions and misconceptions with a view to providing appropriate remedial experiences

Discrimination the ability to distinguish between and among different levels of work or achievement

E-assessment electronic assessment: processes involving the implementation of ICT for the recording, transmission, presentation and processing of assessment material

Edexcel UK examining and awarding body providing a range of qualifications including at higher education level

EiC Excellence in Cities. Government initiative aimed at raising the educational aspirations and attainment of children in inner cities

European Computer Driving Licence European-wide qualification allowing candidates to demonstrate competence in computer skills, covering the areas of basic concepts of IT, using the computer and managing files, word processing, spreadsheets, database, presentation and information, and communication

Formative assessment often called assessment for learning. Assessment used to support teaching and learning, which identifies strengths and weaknesses of the student

GCE General Certificate of Education

GCSE General Certificate of Secondary Education (GCSE). The main secondary school examinations usually at 16, which replaced previous system GCE O-levels and CSEs

GIS Geographic Information System. System of software used for the storage, retrieval, mapping and analysis of spatial data, such as mortality by different regions

GNVQ General National Vocational Qualification. Vocational qualification, often taken as an alternative to GCSE or A-levels, usually after compulsory schooling. Available at three levels; Foundation, Intermediate, and Advanced

High-stakes assessment assessment that has important consequences or implications for students, staff or schools

ICT Information and Communications Technology

Key sills a group of skills valued by employers as being central to all work and learning, including communication, information technology, application of numbers, working with others, and improving own learning and performance

Key Stages the four stages of the National Curriculum: KS1 for pupils aged 5-7; KS2 for 7-11; KS3 for 11-14; KS4 for 14-16

NVQ National Vocational Qualifications. Work-based vocational qualifications. They are portfolio-based qualifications which show skills, knowledge and ability in specific work areas. Can be taken at five levels, depending on level of expertise and responsibility of the job

O-level also GCE Ordinary level. Former system of British examinations taken in a range of subjects, usually at the age of 16. Ran in parallel with but at a higher level than CSE. Both systems now replaced by current GCSE

Parallel forms tests that are created to measure the same constructs, and to produce the same scores, if they were given to individuals on different occasions

PDA Personal Digital Assistant; a small hand-held computer. Depending on level of sophistication may allow e-mail, word processing, music playback, internet access, digital photography or GPS reception

GLOSSARY

Pedagogy philosophy of approach to schooling, learning, and teaching including what is taught, how teaching occurs, and how learning occurs

Portfolio a representative collection of a candidate's work, which is used to demonstrate or exemplify either that a range of criteria has been met, or to showcase the very best that a candidate is capable of

Portfolio assessment assessment based on judgment made about the work shown as evidence within a portfolio

Predictive validity the extent to which scores on a test predict some future performance. For example, a student's GSCE grade can be used to predict their likely A-level grade – in some subjects, the prediction is better than in other subjects

QCA UK public body, sponsored by the Department for Education and Skills (DfES). Roles include the maintenance and development of the national curriculum and associated assessments, tests and examinations

Reliability reliability in measurement and testing is a measure of the accuracy of the score achieved, with respect to the likelihood that the score would be constant if the test were re-taken or the same performance were re-scored by another marker, or if another test from a test bank of ostensibly equivalent items is used

Summative assessment assessment used to measure performance, usually at the end of a course of study

TIMSS Trends in International Mathematics and Science Study, formerly Third International Mathematics and Science Study. Comprehensive study offering data on students' mathematics and science achievement from an international perspective. Data from 1995, 1999, and 2003

UCLES University of Cambridge Local Examinations Syndicate, comprising three business units: Cambridge ESOL (English for Speakers of Other Languages), providing examinations in English as a foreign language and qualifications for language teachers; CIE (University of Cambridge International Examinations), providing international school examinations and international vocational awards; and OCR (Oxford, Cambridge and RSA Examinations), providing general and vocational qualification

Validity the appropriateness of the interpretation and use of the results for any assessment procedure

Value added the increase in learning that occurs during a course of education. Based either on the gains of an individual or a group of students. Requires a baseline measurement for comparison

BIBLIOGRAPHY

Adams, C and Hudson, G (2004). AQA and DRS electronic mark capture, presented at the QCA E-assessment Summit, 24 April

Aim Online P–10 Supplement (2003). Supplement to the VCAA Bulletin No 6 September 2003. AIM Online: www.aimonline.vic.edu.au

Archenhold, WF, Bell, J, Donnelly, J, Johnson, S and Welford, G (1988). Science at Age 15: a Review of APU Findings 1980-1984. London: HMSO

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

Bennett, RE (2002). Inexorable and inevitable: the continuing story of technology and assessment. Journal of Technology, Learning, and Assessment, 1(1). Available from www.jtla.org

Black, P and Wiliam, D (2002). Assessment for Learning: Beyond the Black Box (2002). www.assessment-reform-group.org.uk/publications.html

Chapman, OL and Fiore, MA (2001). Calibrated peer review: a writing and critical thinking instructional tool. The White Paper: a Description of CPR. http://cpr.molsci.ucla.edu/

Cockcroft, WH (1982). Mathematics Counts. London: HMSO

Cohen, Y, Ben-Simon, A and Hovav, M (2003). The effect of specific language features on the complexity of systems for automated essay scoring. Paper presented to the 29th Annual Conference of the International Association for Educational Assessment. www.aqa.org.uk/support/iaea/papers/ben-cohen-hovav.pdf

Cowie, J and Lehnert W (1996). Information extraction. Communications of the ACM vol 39 (1), pp80-91

De Bono, E (2000). Six Thinking Hats. London: Penguin Books

Doiron, G and Isaac JR (2002). Designing an ER online role play for medical students. 2nd Symposium on Teaching and Learning in Higher Education Paradigm Shift in Higher Education, National University of Singapore, 4-6 September 2002

Downes, T and Zammit, K (2000). New literacies for connected learning in global classrooms, in: H Taylor and P Hogenbirk (Eds) Information and Communication Technologies: the School of the Future. London: Kluwer Academic Publishers

EPPI Centre (2002). A Systematic Review of the Impact of Summative Assessment and Tests on Students' Motivation for Learning. http://eppi.ioe.ac.uk

Frederikson, JR and Collins, A (1989). A system approach to educational testing. Educational Researcher, 18(9), 27-32

Freeman, MA and McKenzie, J (2002). Implementing and evaluating SPARK, a confidential web-based template for self and peer assessment of student teamwork: benefits of evaluating across different subjects. British Journal of Educational Technology, 33 (5), pp553-572. Cited at www.educ.dab.uts.edu.au/darrall/sparksite

Getzels, JW and Jackson, PW (1962). Creativity and Intelligence: Explorations with Gifted Students. New York: John Wiley

Kimbell, R (2003). Performance assessment: assessing the inaccessible. Paper presented at Futurelab's Beyond the Exam conference, 19-20 November 2003, Bristol

BIBLIOGRAPHY

Kirriemuir, J and McFarlane, A (2003). Literature Review in Games and Learning (2004). Bristol: Futurelab. Retrieved 05/09/2004 from www.futurelab.org.uk/research/lit_reviews.htm

Klein SP, Hamilton, LS, McCaffrey, DF and Stecher, BM (2000). What do test scores in Texas tell us? RAND Issues Paper. www.rand.org/publications/IP/IP202

Koretz and Barron (1998). The Validity of Gains in Scores on the Kentucky Instructional Results Information System (KIRIS). www.rand.org/publications/MR/MR1014/MR1014.pref.pdf

Linn, RL (2000). Assessments and accountability. ER Online, 29(2). www.aera.net/pubs/er/arts/29-02/linn01.htm

Mathews, JC (1985). Examinations: a Commentary. London: George Allen and Unwin

Messick, S (1995). Validity of psychological assessment. American Psychologist vol 50, no 9, pp741-749

Mitchell, T, Aldridge, N, Williamson, W and Broomhead, P (2003). Computer based testing of medical knowledge. Proceedings of the 7th International Computer Assisted Assessment Conference, Loughborough, pp249-267

Pellegrino, JW, Chudowski, N, Glaser, R (Eds) (2001). Knowing What Students Know. Washington DC: National Academy of Sciences

QCA (2004). The Basic and Key Skills (BKS) E-assessment Experience Report. www.qca.org.uk/adultlearning/downloads/bks_e-assessment_experience.pdf

Richardson, M, Baird, J, Ridgway, J, Ripley, M, Shorrocks-Taylor, D and Swan, M (2002). Challenging minds? Students' perceptions of computer-based World Class Tests of problem solving. Computers and Human Behaviour, 18 (6), 633-649

Ridgway, J and Passey, D (1993). An international view of mathematics assessment - through a class, darkly, in: Niss, M (Ed) Investigations into Assessment in Mathematics Education. Kluwer Academic Publishers, pp57-72

Ridgway, J (1998). The Modelling of Systems and Macro-Systemic Change - Lessons for Evaluation from Epidemiology and Ecology. National Institute for Science Education Monograph 8, University of Wisconsin-Madison

Ridgway, J and Smith, H (2004). Against plagiarism: strategies for defending the validity of assessment systems. EARLI Assessment SIG, Bergen, Norway

Ridgway, J, Swan, M and Burkhardt, H (2001). Assessing mathematical thinking via FLAG, in: D Holton and M Niss (Eds) Teaching and Learning Mathematics at University Level - An ICMI Study. Dordrecht: Kluwer Academic Publishers, pp 423-430. Field-Tested Learning Assessment Guide (FLAG). www.wcer.wisc.edu/nise/cl1

Ridgway J and McCusker, S (2003). Using computers to assess new educational goals. Assessment in Education: Principles, Policy and Practice, vol 10, no 3, pp309-328(20)

Ripley, M (2004). E-assessment question 2004 – QCA keynote speech e-assessment: an overview. Presentation given by Martin Ripley at Delivering E-assessment - a Fair Deal for Learners, a summit held by QCA on 20 April 2004

Roan, M (2003). Computerised assessment: changes in marking UK examinations – are we ready yet? Paper

presented to the 29th Annual Conference of the International Association for Educational Assessment. www.aqa.org.uk/support/iaea/papers/roan.pdf

Robitaille, DF, Schmidt, WH, Raizen, S, McKnight, C, Britton, E and Nicol, C (1993). Curriculum frameworks for mathematics and science. TIMSS Monograph No 1. Vancouver: Pacific Educational Press

Rt Hon Charles Clarke MP, Secretary of State for Education and Skills. Keynote speech at Delivering E-assessment - a Fair Deal for Learners, a summit held by QCA on 20 April 2004

Schulman, L (1998). Teacher portfolios: a theoretical activity, in: N Lyons (Ed) With Portfolio in Hand: Validating the New Teacher Professionalism (pp23-37). NY: Teachers College Press

Slaughter, S and Leslie, LL (1997). Academic Capitalism: Politics, Policies and the Entrepreneurial University. Baltimore: The Johns Hopkins University Press

Sukkarieh, JZ, Pulman, SG and Raikes, N (2003). Auto-marking: using computational linguistics to score short, free text responses. Paper presented to the 29th Annual Conference of the International Association for Educational Assessment. www.aqa.org.uk/support/iaea/papers/sukkarieh-pulman-raikes.pdf

Tattersall, K (2003). Ringing the changes: educational and assessment policies, 1900 to the present, in: Setting the Standard. AQA: Manchester, pp7-27

Teacher Training Agency (2003). Qualifying to Teach: Professional Standards for Qualified Teacher Status and Requirements for Initial Teacher Training

Tomlinson, M (2002). Inquiry into A Level Standards. London: DfES

Tomlinson, M (2004). 14-19 Curriculum and Qualifications Reform: Interim Report Of The Working Group On 14-19 Reform. London: DfES. www.14-19reform.gov.uk

Topping, KJ (1998). Peer assessment between students in college and university. Review of Educational Research. 68 (3), 249-276

APPENDIX: FUNDAMENTALS OF ASSESSMENT

APPENDIX: FUNDAMENTALS OF ASSESSMENT

How shall they be judged?

Here we consider some of the criteria against which tests and testing systems can be judged.

Validity and reliability are often written about as if they were separate things. Actually, they are intimately entwined, but it is worth starting with two simple definitions: validity is concerned with the nature of what is being measured, while reliability is concerned with the quality of the measurement instrument.

A loose set of criteria can be set out under the heading of educational validity (Frederikson and Collins (1989) use the term 'systemic validity'). Educational validity encompasses a number of aspects which are set out below.

Consequential validity: refers to the effects that assessment has on the educational system (Ridgway and Passey (1993) use 'generative validity'). Messick (1995) argues that consequential validity is probably the most important criteria on which to judge an assessment system. For example, high-stakes testing regimes which focus exclusively on timed multiple choice items in a narrow domain can produce severe distortions of the educational process, including rewarding both students and teachers for cheating. Klein, Hamilton, McCaffrey and Stecher (2000), and Koretz and Barron (1998) provide examples where scores on high-stakes State tests rise dramatically over a four-year period, while national tests taken by the same students, which measure the same constructs, show little change.

Construct validity: refers to the extent to which a test measures what it purports to measure. There is a need for a clear description of the whole topic area (the domain definition) covered by the test. There is a need for a clear statement of the design of the test (the test blueprint), with examples in the form of tasks and sample tests. Construct validity requires supporting evidence on the match between the domain definition and the test. Construct validity can be approached in a number of ways. It is important to check on:

- content validity: are items fully representative of the topic being measured?
- convergent validity: given the domain definition, are constructs which should be related to each other actually observed to be related to each other?
- discriminant validity: given the domain definition, are constructs which should not be related to each other actually observed to be unrelated?
- concurrent validity: does the test correlate highly with other tests which supposedly measure the same things?

The essential idea about reliability is that test scores should be a lot better than random numbers. Test situations have lots of reliabilities. The over-arching question concerning reliability is: if we could test identical students on different occasions using the same tests, would we get the same results?

Take the measurement of student height as an example. The concept is easy to define; we have good reason to believe that 'height' can be measured on a single dimension (contrast this with 'athletic ability', or 'creativity' where a number of different components need to be considered). However, the accurate measurement of height needs care.

Height is affected by the circumstances of measurement – students should take off their shoes and hats, and should not slump when they are measured. The measuring instrument is important – a yard stick will provide a crude estimate, good for identifying students who are exceptionally short or exceptionally tall, but not capable of fine discriminations between students; using a tape measure is likely to lead to more measurement error than using a fixed vertical ruler with a bar which rests on each student's head. Time of day should be considered (people are taller in the morning); so should the time between measurements. If we assess the reliability of measurement by comparing measurements on successive occasions, we will under-estimate reliability if the measures are taken too far apart, and students grow different amounts in the intervening period.

Exploration of reliability raises a set of finer-grained questions. Here are some examples:

- is the phenomenon of being measured relatively stable? What inherent variation do we expect? (mood is likely to be less stable than vocabulary size)
- to what extent do different markers assign the same marks as each other to a set of student responses?
- do students of equal ability get the same marks no matter which version of the test they take?

Fitness for purpose: the quality of any design can be judged in terms of its 'fitness for purpose'. Tests are designed for a variety of purposes, and so the criteria for judging a particular test will shift as a function of its intended purpose; the same test may be well suited to one purpose and ill suited to another.

Usability: people using an assessment system – notably students and teachers – need to understand and be sympathetic to its purposes.

Practicality: few designers work in arenas where cost is irrelevant. In educational settings, a major restriction on design is the total cost of the assessment system. The key principle here is that test administration and scoring must be manageable within existing financial resources, and should be cost-effective in the context of the education of students.

Equity: equity issues must be addressed - inequitable tests are (by definition) unfair, illegal, and can have negative social consequences.

About Futurelab

Futurelab is passionate about transforming the way people learn. Tapping into the huge potential offered by digital and other technologies, we are developing innovative learning resources and practices that support new approaches to education for the 21st century.

Working in partnership with industry, policy and practice, Futurelab:

- incubates new ideas, taking them from the lab to the classroom
- offers hard evidence and practical advice to support the design and use of innovative learning tools
- communicates the latest thinking and practice in educational ICT
- provides the space for experimentation and the exchange of ideas between the creative, technology and education sectors.

A not-for-profit organisation, Futurelab is committed to sharing the lessons learnt from our research and development in order to inform positive change to educational policy and practice.

Futurelab
1 Canons Road
Harbourside
Bristol BS1 5UH
United Kingdom

tel +44 (0)117 915 8200
fax +44 (0)117 915 8201
info@futurelab.org.uk

www.futurelab.org.uk

Registered charity 1113051

This publication is available to download from the Futurelab website – www.futurelab.org.uk/research/lit_reviews.htm

Also from Futurelab:

Literature Reviews and Research Reports
Written by leading academics, these publications provide comprehensive surveys of research and practice in a range of different fields.

Handbooks
Drawing on Futurelab's in-house R&D programme as well as projects from around the world, these handbooks offer practical advice and guidance to support the design and development of new approaches to education.

Opening Education Series
Focusing on emergent ideas in education and technology, this series of publications opens up new areas for debate and discussion.

We encourage the use and circulation of the text content of these publications, which are available to download from the Futurelab website – www.futurelab.org.uk/research. For full details of our open access policy, go to www.futurelab.org.uk/open_access.htm.

FUTURELAB SERIES

REPORT 10

ISBN: 0-9544695-8-5