Open-mindedness is an important aspect of CT. Being able to set aside one's own views is a pre-requisite for a fair examination of another's argument. Furthermore, open-mindedness allows a person to acknowledge that their own views may be unsupported or even wrong. Critical Thinking involves a fair assessment of evidence, rather than seeking to support or confirm one's own views.

The definition indicates that CT is a set of skills which one applies not only to other people's reasoning, but also to one's own. Being rational requires analysis, evaluation and elucidation of one's own thinking, with the aim of greater accuracy in one's own reasoning.

Other findings and observations

Mapping of Cambridge Assessment Critical Thinking qualifications and tests

There is only room here for an overview of the mapping findings. In brief, there were, as one might expect, differences in the combinations of sub-skills tested by the various tests, with only one sub-skill common to all, namely 'identifying conclusions'. There was very high congruence between any particular specification and its associated question papers. In just one or two cases, it was judged that some sub-skills were either evidently or implicitly sampled in the question papers or were apparent in the scripts, though not explicit in the specification. It was found that all Critical Thinking products were either substantially or entirely within the definition and taxonomy. Where specifications included sub-skills which were considered not to be Critical Thinking, this was usually attributable to intervention from external agencies.

Skills and Processes which are either on the fringes or more clearly outside the construct of Critical Thinking

Part of understanding what Critical Thinking *is* can be informed by understanding what Critical Thinking *is not*: identifying skills which are frequently confused with Critical Thinking, which lie close to the outer fringes, or may often occur concurrently with genuine Critical Thinking processes. Not all 'higher order thinking' is Critical Thinking.

1. **Reading comprehension**. Whilst reading comprehension is an underlying skill, it is distinct from Critical Thinking. Reading comprehension only asks what is in a passage and may be demonstrated through rephrasing, summarising or précis-ing. Reading comprehension does not, in itself, involve analysing or evaluating. At its closest to Critical Thinking, it involves clarifying the meaning of words or identifying the purpose.
2. **Problem solving**. This uses many reasoning skills and processes which are a facsimile of those in the Critical Thinking taxonomy. The main difference is that the solution to a problem (generally spatial and/or numerical) replaces the argument. Note that here a solution is defined as series of processes leading to the correct answer, and the 'answer' is analogous to a conclusion.The techniques for arriving at a correct solution in problem solving are in many cases different – e.g. trial and error and insight are much more important in problem solving than in Critical Thinking.
3. **Creativity**. An element of creative or imaginative thinking can sometimes be useful in assessing arguments and explanations (thinking up pieces of further evidence or alternative explanations which might undermine the reasoning) and in constructing one's own arguments or taking arguments further. Creativity is not an end in itself and nor is it an essential skill for Critical Thinking. For this reason, it is not contained within the taxonomy.
4. **Sampling issues in evidence**. Size of sample, representativeness, generalisability, understanding the role of a control group – this is all useful knowledge of experimental methods in social science, but in itself is not Critical Thinking. However, such knowledge can be useful to assess credibility and inferences from evidence (e.g. to help identify sweeping generalisations).
5. **Ethical content**, e.g. knowing the names and details of ethical theories, is not part of Critical Thinking. Knowledge of ethical principles, e.g. utililitarianism[6] and deontological theories[7], are on the fringes. Applying such principles and theories to a particular dilemma, however, does involve Critical Thinking.
6. **Syllogism**. This is on the fringes of Critical Thinking. Syllogistic arguments are rarely everyday arguments and, as such, the panel viewed syllogism as an irrelevant technicality for Critical Thinking.

It is hoped that this definition and taxonomy will provide a shared and common understanding of the construct of Critical Thinking. It provides a focus and a fixed reference point for future specification and assessment materials development work. Furthermore, it is hoped this definition and taxonomy will be valuable to teachers and students of Critical Thinking in providing clarity.

References

Black, B. (2007). Critical Thinking – a tangible construct? *Research Matters: A Cambridge Assessment Publication* **2**, 2–4.

Ennis, R.H. (1996). *Critical Thinking*. New York: Prentice Hall.

Facione, P.A. (1990). *Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction. Executive Summary, The Delphi Report*. Millbrae, CA: California Academic Press.

Fisher, A. & Scriven, M. (1997). *Critical Thinking: Its definition and assessment*. Norwich: Centre for Research in Critical Thinking.

Gill, T. & Black, B. (*in prep*). *Do candidates who have taken Critical Thinking AS level perform better in their A levels in other subjects?*

Paul, R. (1992). Critical Thinking: What, Why and How? *New Directions for Community Colleges*, **20**, 1, Spring 1992.

6 The doctrine that the greatest happiness of the greatest number should be the aim of social and political institutions.

7 Ethical theory concerned with rights and duties.

The future of assessment – the next 150 years?

Tim Oates Group Director, Assessment Research and Development

Parts of this article originally appeared in the Spring 2005 bulletin from the Tomorrow Project: 'Shaping the future? Or going with the flow?' They appear here reprinted with the kind permission of the project.

And in today already walks tomorrow Samuel Taylor Coleridge

Prediction is very difficult, particularly if it's about the future Niels Bohr

The paradox is that both of these quotes tap into truths about predicting the future shape of systems. What I will do in this article is look at trends and tendencies in the development of assessment, but also try to offer some theoretical perspectives on why developments take the shape that they do. Bohr is particularly interesting. With startling brevity, he introduces the idea that prediction in natural science is one thing, and in social science, something very different. Assessment systems are lodged in complex, highly interrelated social, political and economic systems. I will initially focus on this issue of what kind of science we can use to predict the future.

What most determines the shape of the future – the sum of individual actions? Ineluctable historical forces? The decisions of a powerful few? Attribution theory has been shown to be a powerful means of exploring why some people make greater progression in life and impact on events than others (Bem, 1972; Lepper *et al*., 1973; Miller *et al*. 1975). Some people feel carried along on a tide of events outside their control, whilst others feel as if they have personal agency – what they do has an effect and they can use this to enhance their world. These different groups attribute the cause of changes in circumstances which affect them to very different things. John Bynner's work at the Centre for Longitudinal Studies, on data from the people in the 1958 and 1970 cohort surveys, has allowed him to develop an insightful notion of 'personal capital' – personal resources upon which people can call to run their lives (Lambe, 2006; Schuller *et al*., 2004). Fundamental to this are feelings of personal power (or powerlessness). The people who display feelings of powerlessness tend to be those with worse outcomes in their lives – encompassing health, education, and social circumstances. His work shows that over time this makes a very substantial difference.

So, a notion that you are being carried along by externally-controlled events is bad for you (and your family). Alongside this, it is interesting to consider how people think of the way that economies, society and history develop. Phrases such as '… the natural operation of competition…', '… the tide of events…', '… the evolution of markets…' and similar crop up time after time in the media. They reinforce the idea of natural processes unfolding through their own unalterable dynamic. And the scale and subtlety of social and economic changes further promote these ideas of events and processes beyond human control – a shift in international markets that brings sudden unemployment to groups of workers and devastates specific communities; subtle changes in family structure brought about by both partners working full-time to sustain family income. Such changes seem far more related to 'natural social and economic evolution' than the results of specific human actions. It is a feeling which is compounded by our wish to attribute responsibility to someone, somewhere (Heider, 1944). This is further reinforced by the difficulty of changing the performance of important social institutions which affect our lives, such as education and health. They are juggernaut in size and structure – substantial investment and policies of direct intervention and change take so much time to bite and take so long to show results to increasingly impatient administrations.

But Realist social theory has re-cast the way we think about the impact of human action on the shape of social systems (Bhaskar, 1975; 1979). Social theories are a part of the social world – they affect the way the social world operates. Roy Bhaskar gives us an excellent example of this important perspective on social theory: the one pound coin. It's a round piece of metal which costs a great deal less than one pound. But it's worth one pound. Why? Because a group of people share a common belief that it's worth a pound. And I'm not knocking this. It is really useful that these shared beliefs operate in the social world. It enables the whole banking system, indeed the whole economy, to work. It shows us that beliefs play an important role in the operation of important social systems.

But while social theory and social research can be very good at explaining things – why certain social groups behave in certain ways – it is also notorious for its lack of precision in predicting events. Natural science is just great at predicting things – like the temperature at which water will boil when I take it up to 6000 metres, or the size of copper wire I will need to safely run a big piece of industrial kit. Frank Achtenhagen has outlined a powerful model of 'planned failure' in social policy (Achtenhagen, 1994). If you fail to adequately understand the nature of the problem you are tackling, you formulate policy which half-engages with the problem, but at the same time putting the policy in place changes the nature of the system you are dealing with, giving you a whole new set of problems which you no longer understand at all. This is the cause of the increasingly-mentioned 'unintended consequences' of policy.

This makes predicting the future a very difficult activity, since the future is partly composed of things which were intended and partly of unintended consequences, and is shaped by the shifting beliefs of people as well as objective forces. Some of these objective forces stem from factors such as limitations on natural resources, others from the impact of policy and action. Runs on banks are fascinating examples of the interplay of subjective and objective forces in social systems. They can be created by crises of confidence – confidence being a subjective human state over which people have individual control – but once people begin to act, based on that personal belief, the crisis becomes an all-too-tangible set of objective forces. They have the economic force and effect of a derailed express train, and appear as something over which individuals can effect little control.

With this as theoretical background I outline some key trends and developments in assessment. I do not advance them as 'the future', but as things which are most likely to feature in the set of factors which shape the future.

There is no shortage of analyses of the inertia in big public systems – interesting analyses of the attempts to reform pension systems, health systems and so on (Bramson and Buss, 2002; Donelan *et al*., 1999; Attwood *et al*., 2003). But the metaphor of 'inertia' does not do justice to the detail of the processes of reform. One new metaphor is needed to describe some of the efforts at change – something which captures a sense of the impetus required to escape the gravitational pull of existing arrangements. This metaphor may be of interest: you can launch a projectile into space on its way to new planets, but if it has inadequate energy, it will fall back to earth, and you end up near where you started, albeit at great expense and with quite a lot of wreckage. This captures the process which currently seems to be occurring in respect of national testing in England: new developments seem to lack the escape velocity to ensure that their purpose, form and operation are genuinely progressive. Innovations seem to be dragged back, by the pull of existing culture, opinion and processes, to a position where they mimic existing arrangements.

The new Single Level Tests were launched by their civil servant authors, in early 2008, as a radical development of national test arrangements (National Assessment Authority, 2008a). Responses to the consultation which followed the launch of the pilot for the tests suggested that the whole model was insufficiently distinctive from current arrangements, and that a range of fundamental measurement issues would prove troublesome in the piloting and operation of the tests. In the first administration of the tests (December 2008), many of these problems were indeed realised. Announcements have now been made (March 2008) regarding a shift in emphasis from using the tests to confirm that learners are 'secure' in a national curriculum level to 'threshold' performance in a level – that is, back to the current focus; and to explore the option of tests covering more than one level (BBC News online, 2008). If these changes are implemented, the supposed radical features of the new arrangements are to be diluted, and the testing arrangements will be far closer to simply providing two sessions, per year, of the existing test model. This brings the risk of testing further dominating the school curriculum (Mansell, 2007) – hardly the intended effect of the original innovation.

This tendency of initiatives to have inadequate 'escape velocity' has been evident in a series of major revisions to the education and training system. It has been particularly evident in vocational education and training. GNVQs are a prime example. Originally conceived with a radical project-based assessment model, GNVQs were constantly modified over a ten year period, each modification bringing the qualification closer and closer to existing assessment approaches in 16–19 general education. This was in part due to an attempt to increase 'parity of esteem' with academic qualifications, but also the result of a power struggle 'for the heart of the qualification' amongst Government agencies. By 2000, as GNVQs became Advanced Vocational Certificates of Education, the qualification had lost many of the features which were associated with learning programmes attractive to the original target group. The qualification had been dragged back to conformance with previous arrangements, no longer fulfilling the role and position which it had been designed for (Oates, 2008). This reduced significantly the range of vocational qualifications capable of being delivered in full time education.

Sometimes the 'pull of gravity' comes not from culture, or the predilections of policy makers, but from deeper structural factors. Although the picture is mixed in terms of quality and patterns of participation, Modern Apprenticeships at level 3 can broadly be considered a success – they are providing a well-grounded practical route to technician level employment. But the numbers of 16–19 years olds participating are startlingly low compared with other European countries which have an apprenticeship route. Total numbers on English apprenticeships at all levels, not just level 3, amount to barely 6% of the cohort, compared to 60% of the cohort in Germany. The causes of this are various, but derive mainly from the state of the labour market. With very low differentials between pay rates during training and pay rates for experienced workers, with training being viewed by hard-pressed employers as a short-term inefficiency, with licence to practice far less established in the UK labour market, and with wage flexibility a cornerstone of increasing employment rates and moving people from welfare to work, the structural conditions and incentive patterns simply militate against mass participation in apprenticeship. Under these conditions, you can try to make the form and content of the learning programmes and qualifications as attractive as possible, but participation simply is not going to undergo any seismic shift.

But whilst innovation is frequently dragged backwards by these processes, there are other societal, economic and technical developments which create constant pressure for change.

First, the explosion in information. The tendencies regarding blurring boundaries between 'private' and 'public' data are clear. In commerce, the patterns of data we leave behind us whilst purchasing goods and services are feeding huge systems of supply management and 'tailored' marketing – the latter presenting loops of feedback which determine in part how we see opportunity and how the commercial world is presented to us. 'Preferences' are recorded when we visit websites…personal profiles of 'you might like this…' built up and played back to us. The formative and summative assessment systems in place and under development fit into this pattern – increasingly fine-grained detail on individual performance, available not only to the learner, but also to teachers and managers of institutions, but also – of course with appropriate safeguards – to the state and its institutions.

In university admissions, in formative assessment in compulsory schooling, in all phases of education and training, there is increasing interest in the detail of performance – unit scores in A and AS examinations, attainment against the individual statements in the National Curriculum, profile components.

The problem here is that we can certainly generate this fine-grained information and we can develop increasingly sophisticated systems to store and display it. Some see the assessment future as being dominated by huge integrated school and college systems which simultaneously hold attendance records, personal data, all school management data (pay, room bookings), learning materials, summative data on individual attainment, formative assessment data, and so on. Apart from the vulnerability and dependency which such systems might stimulate, a key question for assessment is: are we matching our development of such systems with processes by which we can make valid inferences on the basis of these data? Our work with schools on formative assessment tools suggests that teachers do not yet have the skills or techniques to handle these complex arrays of data, and are not yet able to use the data as a basis for differentiated, 'personalised' learning to any great extent.

Richard Kimbell (2007) of Goldsmiths' College, working on the

innovative e-scape assessment project usefully reminds us: '…just because technology allows us to do exciting new things, it doesn't mean that we should do all of them'.

At national policy level, the availability of data on each and every child has led to increasing interest in accountability systems, the data being used as a system management tool within public policy. Many nations considering the future of their assessment arrangements are interested not only in using assessment for school and system monitoring and performance management, but also in international benchmarking – most notably to PISA, PIRLS and TIMSS. This marks a trend of assessment being the hub of control and comparison, as well as supporting more traditional functions associated with learning and progression. This is a heavy weight to carry.

What of the developments 'internal' to assessment? There are interesting things afoot.

The 'empty promise' of adaptive testing?

There was a huge flurry of interest in computer-based adaptive testing in the late 1990s, which waned with the publication of Wainer and Eignor's seminal 2000 review paper (Wainor and Eignor, 2000; Kreitzberg *et al*., 1997). Having expected much from tests which adapt to the performance level of candidates, thus promising greater reliability, reduced test length and/or greater domain coverage, ETS found adaptive systems to be expensive and patterns of item use peculiarly limited within banks – with acute problems of overuse and overexposure of a limited set of items. Expensive, elaborate systems were abandoned and general enthusiasm diminished. The ill-fated on-line KS3 ICT test developed by QCA, funded by the then DfES, started with the intention of having an adaptive model at its heart, but this was quickly abandoned as the complexities hit the development team.

Other issues remain problematic in adaptive tests systems: bank security; comparability problems associated with the facility of a test not being a simple sum of the facility of its items; comparability problems associated with each candidate potentially taking a unique or near-unique combination of items (*op cit*). But small groups of developers have quietly worked away at the provision of working systems – the ESOL group at Cambridge Assessment, Peter Tymms and colleagues at the CEM Centre at Durham, and effective operational systems with robust measurement characteristics are beginning to emerge. Adaptivity may be maturing and emerging as an interesting solution to some of the more enduring problems of mass assessment: the problems of designing single assessments which are accessible to large populations of learners of widely varying levels and patterns of achievement, problems of tiered papers, with their well-known, vicious problems of ill-managed entry/access strategies and equity issues associated with floor and ceiling effects.

On-demand 'test when ready' approaches

'Testing when ready'; 'stage not age', driven by concepts of 'personalised learning' have surfaced as powerful guiding principles for public policy on assessment (BBC News online, 2007). I discuss elsewhere the problems that this may be only superficial rhetoric, with the 'gravity' of existing models and mechanisms pulling innovations back to older, existing models and modes of operation. But these new concepts are nonetheless proving powerful shapers of policy discourse. The new Single Level Tests (SLTs), under pilot in 10 LEAs, are intended to deliver, through six-monthly test opportunities, 'testing when ready' and 'stage not age' assessment. With six-monthly test occasions, candidature in national testing will remain very substantial – with many potentially taking tests more frequently. But even under these conditions, the nature of the entry arrangements pose potential threats to statistically-based standards-maintenance processes. Relatively stable, high population entry is essential to the kind of standards-maintenance processes which are currently used in most educational tests and examinations in England. The potential for only ever having low numbers taking the tests at any given moment (in a fully-blown on-demand system) affects not only the award process but also the ability to see quickly through statistical monitoring any peculiar patterns pointing to defects in the tests/test items. Only having 'when ready' candidates will affect attempts to maintain standards over time, where current fixed test sessions include a mix of 'ready' and 'less ready' candidates. In one legitimate interpretation of 'when ready' testing, an assumption can be made that pass rates should be close to 100% – certainly, the issues of who decides when a person is 'ready', and what the operational definition of 'when ready' actually is, remain problematic.

The drive to 'authentic' tasks

Advocates of ICT-based assessment frequently cite the possibility of setting more complex (aka 'rich', 'dense', 'textured') assessment tasks which assess 'higher order' skills (National Assessment Authority, 2004). This is assumed to be an unmitigated benefit, but the scoring processes, equity issues (in particular the complexities of the tasks and the need for candidates to be clear about what they need to do to succeed in the task), and what constructs are actually being assessed remain highly problematic. An under-recognised issue is that new forms of test may invoke different forms of cognitive engagement. This is illustrated by airline pilot assessment using simulators – you actually want the pilots to believe fully in the test that they are flying – that is, to have full cognitive engagement and no longer be conscious that they are being tested. Should this be emulated, indeed be a goal, in educational testing? There is clear evidence that maintaining awareness of what the test is actually asking for (e.g. seeing past the 'scaffolding') can elevate test performance and can enhance learning. How will tests which emulate the 'simulation' paradigm affect equity (access) for different groups? There can be no simple assumptions that high authenticity, complex tasks should be an ideal in educational testing.

The technological transformation of assessment

Meanwhile, the technological transformation of assessment continues apace, with few commentators doing anything other than picking up on one or two of the full set of ways in which assessment is indeed being transformed:

- Production of assessments (item banking, 'paperless' preparation of 'traditional' exam papers which are then sent direct to printers and then despatched to schools, archiving of materials for reference in comparability studies and standard-setting).

– Provision of on-demand testing, of rapid feedback, and formative assessment.
– Automation of marking of both objective and open response items (automated systems, including those using artificial intelligence – something I deal with below).
– Allocating learners to 'levels' (tiered exam papers replaced by adaptive on-screen tests).
– New ways of presenting questions on screen (development of new types of questions such as those showing rotation of three-dimensional objects, simulations, etc).
– Response by candidates (new types of responses to stimulus material, such as dragging and dropping material).
– Management of scripts (electronic script management – scripts are scanned in and can be sent to markers).
– Restructuring of marking activities (e.g. giving markers the same question from different candidates' papers rather than whole papers to mark).
– Management of results (electronic result management, e.g. texting results to candidates).
– Operation of quality assurance models (e.g. real-time monitoring of markers as they mark on-screen and intervening if problems occur).
– Integration of assessment, learning and MIS information (big school-wide systems).
– Evaluation and research (using scanned scripts and results to run simulations, in order to explore the impact of new assessment processes, but without prejudicing real candidates' chances; integrating assessment data with other data on candidates, such as social background).

Much of the seemingly parochial 'backroom' work on electronic management of question-paper construction, electronic management of scripts (and thus the possibility of new quality assurance processes for marking) is having a huge impact on qualifications. The development of item-level analysis holds huge promise for enhanced quality assurance processes and for research. But the detail of systems matter – not being able go back through marking is a serious weakness in some of the on-screen marking systems; using some forms of scanning prevents markers' annotations from being recorded; … and many of these systems are not so much stable applications as enormous, continuing development projects. But the prize here is almost certainly not reduced cost, but increases in quality and service.

Finally, a few 'emerging issues':

The rise of 'outcomes-based' qualifications in vocational education and training – revised paradigms?

There is a strong international trend towards outcomes-based qualifications (independent of the mode, duration or location of learning) – it is an approach that is reinforced by the commitments intrinsic in the European Qualifications Framework (Oates, 2004; European Commission, 2008). This has the effect of placing high demands on assessment including mastery approaches, high coverage of all necessary skills and knowledge. In addition, it is clear that the concepts of competence embedded in these approaches are crude, and underestimate the importance of vital processes of 'professional formation'. If 'competence-based' models begin to intrude on educational assessment, one of the most important areas to watch will be 'mastery' versus 'compensation' – with mastery tending to demand performance in all elements, thus pointing towards a series of low hurdles rather than items with strongly contrasting facilities/demands. This would represent a fundamental switch in measurement paradigm. Interestingly, this is an important problem facing the policy-makers and developers involved in the Diploma initiative.

Increased enthusiasm for teacher assessment – evidence of benefit in the English setting?

The reviews by Daugherty (Wales) (2004) and Tomlinson (England) (2004) asserted a need to increase the role of teacher assessment in national systems. Neither review presented evidence that teacher assessment can operate in such a way as to deliver stable assessment outcomes in a context of high stakes accountability arrangements. Indeed Sweden offers evidence to the contrary, with acute 'grade inflation' accompanying the introduction of national accountability systems in a system relying heavily on teacher assessment. The principal example of teacher assessment advocated by policy makers etc (Queensland) has not yet integrated accountability arrangements, nor has it generated data on standard reliability measures etc. Classification error is thus difficult to establish – a crucial problem. While the enhancement of learning remains an apparent benefit of such arrangements, the introduction of teacher assessment into a context overdetermined by high stakes accountability arrangements remains highly problematic. What is needed is well-designed research on the technical characteristics of teacher assessment under different system conditions. Without this, a drive towards teacher assessment could well be a leap of faith, in the dark. This carries worrying ethical implications.

Tiering – sufficiently equitable?

Linked to the above, tiering is designed to address clear problems of designing papers which are pitched at the right level for 'bands' of learners (with the specific intention of allowing learners to best demonstrate what they know and can do), but with each specific model for tiering exhibiting undesirable artefacts and deficits. Will tiering continue to be considered by assessment specialists, educationalists, parents and learners as being sufficiently equitable?

Levels (and grades) – are they sustainable?

While the national curriculum legislation requires reporting of children's level of attainment to parents (National Assessment Authority, 2008b), the diagnostic and informative capacity of 'levels' remains under-researched. What do parents make of 'your child is at level 4'? Does it help them direct their support at home in the best way possible (evidence of continued social inequalities in educational outcomes suggest it does not help all families equally)? Levels are blunt; a reduction in diagnostic content in contrast to the scores which make them up – as

Ian Schagen (2003) of NFER has stated in a number of contexts, we spend half our time working scores up into levels and then the rest of the time breaking them back down again to make educational sense of them. Levels introduce a discontinuous scale, with all the attendant problems of two pupils immediately either side of a level boundary being more alike then two pupils at extreme ends of the same level. Misclassification at a key point in a person's educational progression can lead to radically different (inappropriate) educational treatment. The artefacts, identified by QCA's own researchers, around the level thresholds are highly problematic. But many grading systems exhibit similar problems. Both levels and grades may fall foul of increasing public concern for equitable treatment in both access to learning and in educational measurement.

Attacks on the possibility of maintenance of standards over time – and intolerance of measurement error

The gap between public understanding of assessment and expectations of technical rigour remains wide (Wood, 1993; Newton, 2005). There is increasing commitment of assessment specialists and managers to enhance public understanding, with the pressure this brings for more realistic expectations regarding the difficulty – and indeed the sense – of maintaining standards over anything but short time frames. Concern over maintaining standards may be increasingly replaced by concerns to ensure that qualifications are fit for purpose in respect of ever-changing societal, labour market and economic requirements.

Whilst this article may not offer the apparent certainties peddled by futurologists (warning: believe them and that might make it true), it tries to map out some of the trends and tendencies which are playing a part in shaping the future. Perhaps the most important message from this analysis is that our intentions DO matter – the values which we hold will shape events and systems. Clarity of purpose and a firm accountability to learners would seem to be a vital bedrock under the shifting sands of public assessment systems.

References

Achtenhagen, F. (1994). *How should research on vocational and professional education react to new challenges in life and in the worksite*. Paper presented to International Research Network on Training and Development (IRNETD) Conference, Milan, March 1994.

Attwood, M., Pedlar, M., Pritchard, S., & Wilkinson, D. (2003). *Leading change – a guide to whole system working*. Bristol: Polity Press.

BBC News online (2007). *'Testing when ready' gets going* http://news.bbc.co.uk/1/hi/education/7137149.stm Accessed 11 04 08.

BBC News online. (2008). *Pilot progress tests made easier*. http://news.bbc.co.uk/1/hi/education/7246871.stm Accessed 11 04 08.

Bem, D. (1972). Self-perception theory. In: L. Berkowitz (Ed.), *Advances in experimental social psychology*. Vol. 6. New York: Academic Press.

Bhaskar, R. (1975). *A realist theory of science*. Falmer: Harvester.

Bhaskar, R. (1979). *The possibility of naturalism: a philosophical critique of the contemporary human sciences*. London: Harvester.

Bramson, R. & Buss, T. (2002). Methods for Whole System Change in Public Organizations and Communities: An Overview of the Issues. *Public Organisation Review*, **2**, 3, 211–221.

Daugherty, R. (2004). *Learning pathways through statutory assessment: key stages 2 and 3. Interim report of the Daugherty Assessment Review Group*. Daugherty Assessment Review Group, Cardiff.

Donelan, K., Blendon, R.J., Schoen, C,, Davis, K., & Binns, K. (1999). The cost of health system change: public discontent in five nations. Harvard Opinion Research Program, Harvard School of Public Health, Boston, USA. *Health Affairs*, **18**, 3, 206–16.

European Commission. (2008). http://www.europe.org.uk/news/view/-/id/218/Accessed 11 04 08.

Heider, F. (1944). Social perception and phenomenal causality. *Psychological Review*, **51**, 358–374.

Kimbell, R. (2007). *Technology and the assessment of creative performance*. Technology Education Research Unit, Goldsmiths College, London. Keynote presentation at Cambridge Assessment Conference, 15th Oct 2007.

Kreitzberg, C. B., Stocking, M. L., & Swanson, L. (1997). *Computerized Adaptive Testing: The Concept and Its Potentials*. Princeton: Educational Testing Services.

Lambe, B. (2006). *Conceptualising and measuring agency using the British Household Panel Survey data*. BERA annual conference 2006, University of Warwick.

Lepper, M., Greene, D., & Nisbett, R. (1973). Undermining children's intrinsic interest with extrinsic reward: A test of the "overjustification" hypothesis. *Journal of Personality and Social Psychology*, **28**, 129–137.

Mansell, W. (2007). *Education by numbers: the tyranny of testing*. London: Politico's Publishing.

Miller, R., Brickman, P., & Bolen, D. (1975). Attribution versus persuasion as a means of modifying behaviour. *Journal of Personality and Social Psychology*, **31**, 430–441.

National Assessment Authority. (2004) http://www.naa.org.uk/naaks3/documents/2004_KS3_ICT_report.pdf Accessed 11 04 08.

National Assessment Authority. (2008a). http://www.naa.org.uk/single_level_tests/ Accessed 11 04 08.

National Assessment Authority. (2008b). http://www.qca.org.uk/eara/ Accessed 11 04 08.

Newton, P. (2005). The public understanding of measurement inaccuracy. *British Educational Research Association Journal*, **31**, 419–442.

Oates, T. (2004). The role of outcomes-based qualifications in the development of an effective vocational education and training (VET) system. *Policy Futures in Education*, ISSN 1478–2103, **2**, 1.

Oates, T. (2008). Going round in circles: temporal discontinuity as a gross impediment to effective innovation in education and training. *Cambridge Journal of Education*, **38**, 1.

Schagen, I. & Hutchison, D. (2003). Adding value in educational research – the marriage of data and analytical power. *British Educational Research Journal*, **29**, 749–765.

Schuller, T., Bynner, J. & Feinstein, L. (2004). *Capitals and capabilities*. Centre for Research on the Wider Benefits of Learning.

Tomlinson, M. (2004). *14–19 Qualifications and curriculum reform*. Nottingham: DfES.

Wainer, H. & Eignor, D. (2000). Caveats, pitfalls and unexpected caveats of implementing large-scale computerised testing. In H. Wainer (Ed.), *Computerized adaptive testing: a primer*. Chapter 10. 2nd edition. New Jersey: Lawrence Erlbaum Assoc Inc.

Wood, R. (1993). *Assessment and testing*. Cambridge: University of Cambridge Local Examinations Syndicate.