INTRODUCTION

In recent years educators and policymakers in higher education have expressed a preference for assessment which:

- evidences thinking and problem-solving skills rather than discrete knowledge (Berlak et al, 1992; CSUP, 1992; Dearing, 1997; Taylor, 1994);
- 'directly informs instruction' (Nichols, 1994);
- represents meaningful, significant and worthwhile forms of human endeavour and accomplishment (Wiggins, 1989a).

These preferred forms of assessment, generically referred to as alternative assessment (Birenbaum, 1996) have developed in part from recent evidence on learning. Contemporary cognitive psychology indicates that meaningful learning is reflective, constructive and self-regulated and that all learning requires learners to think and actively construct evolving mental models. Learning is now understood to proceed in many directions at once and at an uneven pace. The realisation of what is involved in meaningful learning together with the recognition of the role of social context in shaping higher-order cognitive abilities and dispositions suggests that what is important in assessment is evidence of how and whether students organise, structure, and use information in context to solve complex problems (Baxter & Glaser, 1998). Advocates argue that this alternative form of assessment will help students to develop the conceptual and analytical skills needed to prepare them for future vocational success (Garcia & Pearson, 1994). However, while the lobby for alternative assessment practices appears to have considerable support, the range of terms, such as performance assessment, authentic assessment, direct assessment, constructive assessment, embedded assessment (Baker et al, 1993; Berlak et al, 1992; Biggs, 1999; Hakel, 1998; Race, 1999; Wiggins, 1989a; 1989b; 1993; Yorke, 1998), that have been posited as examples or variants of alternative assessment suggests that the construct is,

as yet, insufficiently robust to be used with any degree of consensus. Given the complex academic, functional and social purposes of higher education (Bowden & Marton, 1998; Dearing 1997) and given the alleged claims for alternative assessment to promote self-motivated, self regulated and successful learners, this paper seeks to gain some clarity on the construct of alternative assessment through a conceptual analysis of its validity in higher education.

ALTERNATIVE ASSESSMENT AND THE NEED FOR ACCOUNTABILITY

Broadly speaking alternative assessment is characterised as an alternative to standardised, norm-referenced, multiple-choice testing and typically claims the following features (Linn et al, 1991; Linn & Baker, 1996; Wiggins, 1989b):

- student involvement in setting goals and criteria for assessment;
- performing a task, creating an artefact/product;
- use of higher-level thinking and/or problem solving skills;
- measuring metacognitive, collaborative and intrapersonal skills as well as intellectual products;
- measuring meaningful instructional activities;
- contextualisation in real-world applications;
- use of specified criteria, known in advance, which define standards for good performance.

Alternative assessment would thus reject the fairly fundamental beliefs that have informed traditional assessment: that there can be universality of meaning as to what any grade or score represents; that it is possible to separate the goals of education from the means for their attainment; and that it is possible to conceive of learning as *either* cognitive *or* affective *or* conative (Berlak et al, 1992). Instead, alternative assessment implies that there need to be new formats for gathering information about students'

achievements, that there have to be new processes through which such information is synthesised (in order to determine/diagnose achievement) and that the formats and processes should seek to serve the welfare of each student. That the primary beneficiary of assessment should be the learner or student is repeatedly asserted in the literature (for example, Biggs, 1999; Black & Wiliam, 1998; Bowden, & Marton, 1998; Crooks, 1988). In other words, assessment is viewed as having a primarily formative function.

Whether or not alternative assessment can succeed in eliminating the negative test-preparation effects of multiple choice testing, have a positive influence on learning and instruction, measure higher order skills and motivate students as its proponents would claim (Berlak et al, 1992; Garcia & Pearson, 1994; Wiggins, 1989a; 1989b; 1993) is still being substantiated (Taras, 2002). However, even with confirming evidence, it would be unwise to assume that alternative assessment is the panacea for all assessment problems in higher education. It cannot be denied that assessment in higher education is, to some extent, a high stakes enterprise. In other words it can have critical consequences. Insofar as the current political climate is one of high accountability for how the public purse is being spent, there is clear expression of interest by the UK government to learn about, and publicise, the extent of student achievement (Filer, 2000). However well or badly the data on such achievement are communicated and publicised, the principal means through which they are gathered is that of assessment. Further, within the higher education sector, the preparation of our graduates to take their professional place in society necessitates benchmarks through which student entry, progress, qualification and graduation is recorded (Cam, 2001; Sutherland, 2001). It is commonly recognised that the mechanism through which such certification operates is that of assessment. So, however much we might want to be primarily concerned with the diagnosis and support of student learning, the reality is that assessment in higher education is not confined to instructional improvement. The

currently dominant view that higher education is one means of improving the country's social and economic goals means that higher education, at least in part, responds to particular political agenda and so, while the use of assessment for administrative purposes is not unwarranted, this particular use can acquire a primary rather than secondary purpose (Evans, 2002; Harvey & Knight, 1996; Wolf, 2002). In other words the external pressures on higher education may cause assessment to assume a primarily summative function. Because assessment is viewed by policymakers as an agent of educational reform (Linn, 2000), comparisons and generalisations on the basis of derived data are a logical consequence. If alternative assessment is providing the data that inform educational policy, the extent to which alternative assessment is valid has to be of central concern.

THE VALIDITY OF ALTERNATIVE ASSESSMENT

The validity of any assessment is its most important quality (Crooks et al, 1996). Early conceptions of validity as the extent to which an assessment instrument measures whatever it purports to measure have been refined into a "summary of both the existing evidence for and the potential consequences of score interpretation and use" (Messick, 1989; p. 13). In other words what is to be deemed valid is not the assessment instrument used or the resulting scores per se but the *inferences* which are derived from either. Although, as a consequence of historical practice, validity can be characterised as different types, Messick (1989) argues that validity is a unified concept (albeit differentiable into distinct aspects) which is best represented in the term, construct validity. The pivotal role of construct validity in evaluating alternative assessment is perfectly consistent with the aims of higher education to have students develop the cognitive abilities of thinking, reasoning, planning and decision-making in the service of genuine problem solving. In trying to determine what assessment results signify and how they help us to understand an individual, we are summarising or accounting for

the regularities or relationships in observed behaviour and thereby *constructing* our inferences. Two aspects of construct validity - task specification and consistency of marking - appear to be particularly relevant when considering high-stakes, alternative assessment in higher education, since it is the nature of the assessment task together with how the learner's performance on the task is judged that jointly indicate what learning is deemed important (Bereiter & Scardamalia, 1987; Boud, 1990; Berlak et al, 1992; Khattri et al, 1998).

Task Specification

Most assessment assumes that indicative tasks are merely samples of the domain that is being assessed which, in turn, necessitates that the tasks are representative of the domain in question. In other words the tasks should demand the procedural, conditional and declarative knowledge (Alexander et al, 1991; Dole et al 1991; Paris et al, 1983) required for mastery of the specified domain. If construct representation is compromised the assessment task may be too narrow resulting in construct under-representation or may be too broad resulting in construct-irrelevance (Messick, 1989). While both of these validity threats are pertinent to all educational assessment, they pose a fairly significant problem for alternative assessment.

Because alternative assessment is concerned with complex multi-faceted performances/products, because alternative assessment allows student choice and negotiation, because alternative assessment can find manifestation in range of heterogeneous devices, it is not difficult for irrelevant variables to be used in making judgements about achievement. The most common construct irrelevant variables are ancillary skills and knowledge which can contaminate inferences about assessment performance, particularly when there is no clear distinction between the purpose of the assessment and the skills needed to respond correctly to the assessment task (Wiley, 1990). So, for example, if the purpose of the assessment was to judge how well the

student teacher could motivate a class of pupils, then observing the student teacher actually managing a class might be a relevant means of assessing. However, one would be introducing an irrelevant variable through trying to achieve this same assessment purpose by requiring the student to write an essay on motivation. But if the purpose of the assessment was to judge how well the student could analyse and evaluate different perspectives on motivation, writing a critical essay would probably not be an ancillary demand. What this example highlights is that skills which may be ancillary for one interpretation of assessment performance may be relevant for another interpretation. It is similarly possible to overlook the place ancillary knowledge may play in making judgements and thereby discriminate against ethnic diversity and contribute to inequitable assessment practice (Baker & O'Neil, 1994; Sackett et al, 2001). If, further, multiple and alternative devices are permissible in the assessment of the construct, it would not be surprising if the potential contaminating effects of ancillary skill and knowledge were to increase. While within the philosophy of alternative assessment there is no reason to expect performance on one task to be similar to that on another, or to expect that assessment rubrics be standardised (since this very standardisation could preclude the assessment of important skills such as conceptualising a problem), the possibility of construct irrelevance contamination complicates the extent to which performance on 'equivalent' tasks can be considered comparable. Some might want to dispense with this concern on the basis of the argument that assessment can be valid without necessarily conforming to psychometric notions of reliability (Linn & Baker, 1996; Moss, 1992; 1994). However, Moss's (1994) argument was premised on the view that while information from multiple tasks could improve the validity of the judgement made, there had to be "consistency among independent measures intended as interchangeable" (p. 10). This qualification does not therefore appear to give carte blanche to proponents of alternative assessment. Since, as was argued above, assessment in higher education has high stakes for students, the principle of equity demands that attention be given to rigorous task specification in

order that irrelevant variables are not contaminating results.

A further problem for alternative assessment is the possibility of attenuating, or even discounting, construct representation. The requirement for learners to demonstrate that they have mastered specific skills and competencies by doing something or producing something means that the constituent task-specific skills of some performance may well constitute the evaluation criteria (Motowidlo et al, 1990; Russell & Kuhnert, 1992). If all that counts is the quality of the artefact or performance offered for evaluation, then task-specific assessment can be perfectly adequate. So long as the assessment task elicits the skills underlying the performance in the domain of interest (as in acting, dancing, painting, participative sport and so on) there can be little quibble about the validity of the task. That the performance per se and the target of assessment are essentially the same thing is what Messick (1994) refers to as task-driven performance assessment. However, it does not follow that task-driven performance is always appropriate. Higher education is rarely concerned with one particular performance. If people are to learn to think, reason, plan and make good decisions (which is a significant aim of higher education), they must be able to generalise what they have learned in the past to new learning and be able to apply and extend their learning to a range of situations (Haskell, 2001). Because of this need to generalise abstract concepts (Bereiter, 2002) from one situation to another, task-driven performance should not be a significant part of educational assessment. Rather, the concern to assess whether or not a person understands the underlying attributes or variables which represent the crucial components of the skilled performance (and thus draw on them at will) means that the performance assessment should be what Messick (1994) terms construct-driven (in which the knowledge, skills or other attributes to be assessed guide the selection of the task as well as the development of the scoring procedures). Although it is argued here that construct-driven assessment is preferable - because in task-driven assessment generalisable learning, which is necessary for high-

stakes assessment, can get lost (Schavelson et al, 1992) - not everyone in higher education will necessarily agree, resulting in major confusion when one interpretation of performance assessment rather than another is assumed, particularly when presumptions often determine the type of evidence deemed sufficient for validity (Messick ,1989).

It is, however, possible to examine whether particular performance tasks are functioning as intended (Messick ,1989). Through using protocol analyses (in which participants 'think aloud' during or after problem solution), analysis of reasons (in which participants provide a rationale for their responses) or analysis of errors (in which assessors draws inferences from participants' incorrect representation or implementation of the problem), it is possible to examine how well the performance task as conceived by the assessors actually measures complex cognitive processes. Through analysing verbal protocols Baxter & Glaser (1998) found tasks to range from those that required in-depth understanding of subject matter knowledge to those that relied only on the information given in the assessment task. Similarly they found that tasks ranged form being very open to very constrained in the way that content knowledge and process skill could be combined to complete the performance. In another study Hamilton et al (1997) found that when the task was so open-ended as to merely *imply* the cognitive processing intended, students did not necessarily use their resources but, rather, relied on common sense reasoning. On the other hand when the task was more structured, students were more focused on scientific reasoning. It cannot therefore be assumed that performance assessments will demand greater cognitive complexity because to the assessor(s) they appear to do so. Given that performances can vary in what underlying cognitive processing they reveal, the lack of clear and comprehensive task specification for successful performance would appear to be a source of invalidity in, and therefore a difficulty for, alternative assessment. However, even if task specification were to be panoptic and unequivocal, it would

nevertheless remain an interpretation of the domain in question, since there can always be debate about the behavioural manifestations of abstract, psychological constructs such as higher order skills, problem solving and critical thinking (Baker et al, 1993). One possible consequence of this is that what is assessed is seen as more important than what is not (Messick, 1989). Such distortion of values associated with the domain may in turn result in learning being conceived as instrumentally rather than intrinsically important; a tension that is well recognised but not easily resolved (Linn, 2000). While the ideology of alternative assessment would imply a resolution of this tension, the need for high stakes assessment to be reliable renders alternative assessment in higher education problematic, as will be further elucidated in the ensuing discussion.

Consistency of Marking

The move towards alternative assessment is premised on the view that cognitive and situative perspectives best explain how complex learning occurs. Within these perspectives, assessment is not only about judging how much people know but judging how, when and whether they use what they know. Because of this emphasis on higher-order processing, alternative assessment is concerned to assess the products and processes of cognitive and social functioning. Students therefore can have considerable latitude in interpreting the stimulus task and constructing their responses in alternative assessment, which makes for difficulty in the reliable interpretation the performance. Historically, the dominant method of interpreting performance in educational assessment has been by comparing the results of one individual with those of a well-defined reference group. The data from the relevant reference group contextualise the extent to which the individual's performance is consistent with/deviant from average. While such norm-referencing usefully gives meaning to measures such as blood pressure or cholesterol level, it is arguably less useful in

educational assessment because it does not describe students' actual achievements (Glaser, 1963; 1990). To redress this perceived deficiency, predetermined levels or standards of performance became the basis for comparison in order to be able to provide explicit information as to what students can and cannot do. There are, however, some difficulties with referencing interpretations of performance in terms of criteria.

One is in determining the criteria. Because of the need to provide explicit information as to what students can do, the specification of what elements of performance are desired and what the criteria of excellent and adequate performance are in each case (Resnick & Resnick, 1993) can become precise and elaborate. A potential disadvantage in such detailed specification is that the assessment task is reduced to a set of routine, algorithmic subtasks making no authentic demands of the student, and thereby negating the pedagogical and philosophical underpinnings of alternative assessment (Wiliam, 1998). Confusingly, the converse also obtains. While there is some (though not unanimous) consensus within the academic community as to, for example, what an essay might look like (Hounsell, 1997), the definition of other critical performances continues to evolve (Moss, 1992). That definitive standards are almost altogether lacking in education can therefore lead to ambiguity and variability of practice in the determination of criteria. Among the criteria that should be included, according to Linn et al (1991), are cognitive complexity (the processes of higher order thinking that are required to be exercised), content quality (the depth of subject matter expertise) and content coverage (the breadth of domain representation); constructs which are covert and therefore non trivial to either conceptualise or represent. Given the potential difficulty in task specification rehearsed above, it is not difficult to appreciate that the determination of assessment criteria might be problematic.

Another difficulty is in using the criteria. Because the whole point of alternative assessment is not to award a single score or percentile rank, but to judge a multi-faceted accomplishment, the issue of human judgement becomes significant. And since human judgement about any particular event can differ, dramatically, both within persons across time and amongst persons, the reliability of alternative assessment is a serious issue. When judgements about the same event differ, whose judgement should be the benchmark? Because "judges should know specifically where in performance to look and what to look for" (Wiggins, 1992) the issue of reliability is often seen as being resolved in the specification of clear criteria. However, as Wiliam (1996a) points out, consistency does not reside in external, pre-specified criteria and so to believe that reliable marking is a function of specifying clear criteria is naïve. That criteria themselves are the subject of interpretation is recognised in the practices of training and moderation where individuals learn to rate performances to agreed standards or otherwise acquire shared understanding of performance standards (Baker et al 1993; Resnick & Resnick, 1993). In the process of rating, one's substantive knowledge, one's contextually derived expectations of what is appropriate and one's beliefs as to how learning occurs all subtly influence, and thereby mediate, the judgements made (Baker & O'Neill, 1994). In other words, as Angoff (1974) pointed out many years ago, "lurking behind the criterion-referenced evaluation, perhaps even responsible for it, is the norm-referenced evaluation" (p. 4). The espoused need for training in reliable rating is clear evidence that consistency in marking is achieved through the shared values, meanings and understandings of the markers that must originally derive from normative assumptions; leading inexorably to the conclusion that criterion-referenced assessment is not as distinct from norm-referenced assessment as we might like to believe. Because the determination and use of criteria for performance assessment are not unproblematic, the validity of alternative assessment continues to be a matter of concern.

If alternative assessment continues to be important in determining achievement in higher education, it is right and proper that there also be further work and deliberation to resolve the attending validity issues. In the wake of criterion-referencing being understood as inherently problematic, Wiliam (1996b; 1998) proposes that construct-referenced assessment would better fulfil the aims of alternative assessment. Construct-referenced assessment assumes that what it means to be competent in a particular domain is well conceptualised amongst experts/practitioners in the domain and so while experts/practitioners might not agree on definitions of performance/specific criteria to be adhered to in any assessment task, they would agree, at least tacitly, on possible examples of appropriate/inappropriate behaviour to represent the construct(s) under consideration (Wiliam, 1998). Construct-referenced assessment would thus be consistent with constructivist perspectives on learning that stress the essentially social and situated nature of human cognition (Kirshner & Whitson, 1997). While construct-referenced assessment is not yet commonly applied in higher education, neither is it unknown. The examination of the PhD thesis is an example of construct-referenced assessment in which the decision to award the degree is made by one or more experts. More fundamentally, however, the examiners are neither judging achievement, nor predicting future performance but are instead inaugurating individual entry into a community of practice. These common understandings within any community of practice do not evolve naturally but are constructed out of dissent and reasoned argument to further the process of enquiry (Kirshner & Whitson, 1997). Wiliam's (1998) argument for construct-referencing is persuasive in the optimism it holds for the formative function of alternative assessment but whether or not construct-referencing will ever become common practice seems in large measure to be constrained by the summative function of assessment in higher education. For as long as we have to describe and differentiate between the achievement of individual, and cohorts of, students we are involved in a process of measurement. This is inherently a flawed process which, in the interests of all

embroiled in summative assessment, must be as transparent as possible to attenuate the potential contamination of the many sources of measurement error. Alternative assessment would not seem to be an immediately convincing form of assessment within the current assessment realities and constraints.

CONCLUSION

Advances in our understanding of learning have partly influenced the inception and use of alternative assessment in higher education. While alternative assessment devices take a variety of forms, they essentially privilege the students' own conceptualisations of their experiences. The dominance given to students' interpretations of their world is well suited to formative assessment which is concerned with the facilitation of learning. However, higher education must also be concerned with summative assessment for reasons of accountability and certification. The extent, therefore, to which alternative assessment is valid, must be considered. The bulk of the extant literature would suggest that task specification in alternative assessment is problematic because of the unwitting ease with which construct irrelevance and construct under-representation can contaminate the assessment devices. The literature would also suggest that marker consistency is problematic because both how assessment performance is to be interpreted and the reliability with which persons can make interpretations is very variously understood. Because of these difficulties it would be cautious to conclude that while alternative assessment may be instructionally informative, its use for summative and accountability purposes is much less prudent. To this extent, alternative assessment in higher education it is not a particularly convincing form for high-stakes assessment.

REFERENCES

Alexander, P., Schallert, D. and Hare, V. (1991) Coming to terms: how researchers in

learning and literacy talk about knowledge, *Review of Educational Research*, 61, 3 pp.315-43.

Angoff, W. (1974) Criterion referencing, norm referencing and the SAT, *College Board Review,* 92, pp.3-5, 21.

Baker, E., O'Neil, H. &Linn, R. (1993) Policy and validity prospects for performance-based assessment, *American Psychologist* 48, pp.1210-8.

Baker, E., & O'Neil, H. (1994) Performance assessment and equity, *Assessment in Education* 1(1), pp.11-26.

Baxter, G. & Glaser, R. (1998) Investigating the cognitive complexity of science assessment, *Educational Measurement: Research and Practice*, 17, 3, pp.37-45.

Bereiter, C. & Scardamalia, M. (1987) *The Psychology of Written Composition* (NJ, Lawrence Erlbaum Associates).

Bereiter, C. (2002) *Education and Mind in the Knowledge Age* (NJ, Lawrence Erlbaum Associates).

Berlak, H., Newmann, F, Adams, E., Archbald, D., Burgess, T., Raven, J. & Romberg, T. (1992) *Towards a New Science of Educational Testing and Assessment* (New York State, University of New York Press).

Biggs, J. (1999) *Teaching for Quality Learning at University* (Buckingham, The Society for Research into Higher Education &The Open University Press).

Birenbaum, M. (1996) Assessment 2000: towards a pluralistic approach to assessment, in: M. Birenbaum & F. Dochy (Eds.) *Alternatives in Assessment of Achievements, Learning processes and Prior Knowledge* (Dordrecht, Kluwer Academic Press).

Black, P. & Wiliam, D. (1998) Assessment and classroom learning, *Assessment in Education* 5(1), pp.7-74.

Boud, D. (1990) Assessment and the promotion of academic values, *Studies in Higher Education* 15(1), pp.101-111.

Bowden, J. & Marton, F. (1998) *The University of Learning* (London, Kogan Page).

Cam, P. (2001)The French Baccalauréat since 1985: level of qualification or type of diploma, *Assessment in Education*, 8, 3, pp.291-314.

Committee of Scottish University Principals (1992) *Teaching and Learning in an Expanding Higher Education System* [The MacFarlane Report] (Edinburgh, CSUP).

Crooks, T. (1988) The impact of classroom evaluation practices on students, *Review of Educational Research* 58 (4), pp.438-81.

Crooks, T., Kane, M. & Cohen, A. (1996) Threats to the valid use of assessment, *Assessment in Education*, 3, 3, pp.265-285.

Dearing, R. (1997) *National Committee of Inquiry into Higher Education (Dearing Report),* Higher Education in the Learning Society, Report of the National Committee (Norwich, HMSO).

Dole, J., Duffy, G., Roehler, L. and Pearson, P. (1991) Moving from the old to the new: research on reading comprehension instruction, *Review of Educational Research*, 61, 2, pp.239-64.

Evans, G. (2002) *Academics and the Real World* (Buckingham, The Society for Research into Higher Education &The Open University Press).

Filer, A. (2000) *Assessment: social practice and social product* (London, Routledge).

Garcia, G. & Perarson, P. (1994) Assessment and diversity. *Review of Research in Education*, 20, pp.337-391.

Glaser, R. (1963) Instructional technology and the measurement of learning outcomes: some questions, *American Psychologist*, 18, pp.519-21.

Glaser, R. (1990) Toward new models for assessment, *International Journal of Educational Research*, 14(5), pp.475-83

Hakel, M. (1998) *Beyond Multiple Choice* (Mahwah, Lawrence Erlbaum Associates).

Hamilton, L., Nussbaum, E., & Snow, R. (1997) Interview procedures for validating science assessments, *Applied Measurement in Education*, 10, 2, pp.181-200.

Harvey, L. & Knight, P. (1996) *Transforming Higher Education* (Buckingham, The Society for Research into Higher Education & The Open University Press).

Haskell, R. (2001) *Transfer of Learning* (London, Academic Press).

Hounsell, D. (1997) Contrasting conceptions of essay writing, in: F. Marton, D. Hounsell & N. Entwistle (Eds) *The Experience of Learning* (Edinburgh, Scottish Academic Press).

Khattri, N., Reeve, A. & Kane M. (1998) *Principles and Practices of Performance Assessment* (NJ, Lawrence Erlbaum Associates).

Kirshner, D. & Whitson, J. (Eds.) (1997) *Situated Cognition* (Mahwah, Lawrence Erlbaum Associates).

Linn, R., Baker, E. & Dunbar, S. (1991) Complex performance-based assessment: expectations and validation criteria, *Educational* Researcher 20(8), pp.15-21.

Linn, R. & Baker, E. (1996) Can performance-based student assessments be psychometrically sound? In: J. Baron & D. Wolf (Eds.) *Performance-Based Student Assessment: Challenges and Possibilities* (Chicago: Ninety-Fifth Yearbook of the National Society for the Study of Education).

Linn, R. (2000), Assessments and accountability, *Educational Researcher*, 29(2), pp.4-16.

Messick, S. (1989) Validity, in R.Linn (Ed) Educational Measurement, 3rd ed. (pp. 13-103). New York: Macmillan.

Messick, S. (1994) The interplay of evidence and consequences in the validation of performance assessments, *Educational Researcher* 23(2), pp.13-23.

Moss, P. (1992) Shifting conceptions of validity in educational measurement: implications for performance assessment, *Review of Educational Research* 62, pp.229-58.

Moss, P. (1994) Can there be validity without reliability? *Educational Researcher*, 23, pp.5-12.

Motowidlo, S., Dunnette, M. & Carter, G. (1990) An alternative selection procedure: the low-fidelity simulation, *Journal of Applied Psychology* 75, pp.640-7.

Nichols, P. (1994) A framework for developing cognitively diagnostic assessments. *Review of Educational Research*, 64(4), pp.575-603.

Paris, S., Lipson, M. and Wixson, K. (1983) Becoming a strategic reader, *Contemporary Educational Psychology*, 8, pp.293-316.

Race, P. (1999) Why assess innovatively?', in: S. Brown & A Glasner (Eds) *Assessment Matters in Higher Education* (Buckingham, The Society for Research into Higher Education &The Open University Press).

Resnick, L. & Resnick, D. (1993) Assessing the thinking curriculum:.new tools for educational reform, in: B. Gifford & M. O'Connor (Eds.) *Changing Assessments: alternative views of aptitude, achievement and instruction.*(The Netherlands, Kluwer Academic Publishers).

Russell, C. & Kuhnert, K. (1992) New frontiers in management selection systems: where measurement technologies and theories collide, *Leadership Quarterly* 3, pp.109-36.

Sackett, P., Schmitt, N., Ellingso, J. & Kabin, M. (2001) High stakes testing in employment, credentialing and higher education, *American Psychologist* 56(4), pp.302-18.

Schavelson, R., Baxter, G. & Pine, J. (1992) Performance assessments: political rhetoric and measurement reality, *Educational Researcher* 21(4), pp.22-7.

Sutherland, G (2001) Examinations and the construction of professional identity: a case study of England 1800-1950, *Assessment in Education*, 8(1), pp.51-64.

Taras, M. (2002) Using assessment for learning and learning from assessment, Assessment & Evaluation in Higher Education, 27(6), pp.501-510.

Taylor, C. (1994) Assessment for measurement or standards: the peril and promise of large-scale assessment reform, *American Educational Research Journal*, 31(2), pp. 231-62.

Wiggins, G. (1989a) Teaching to the (authentic) test, *Educational Leadership* 46(7), pp.41-47.

Wiggins, G. (1989b) A true test: toward more authentic and equitable assessment, *Phi Delta Kappan* 70, pp.703-713.

Wiggins, G. (1992) Creating tests worth taking, *Educational Leadership* 49(8), pp.26-33.

Wiggins, G. (1993) *Assessing Student Performance* (San Francisco, Jossey-Bass).

Wiggins, G. (1998) An exchange of views on semantics, psychometrics and assessment reform: a close look at authentic assessments, *Educational Researcher* 55, pp.19-22.

Wiley, D. (1990) Test validity and invalidity reconsidered, in: R. Snow & D. Wiley (Eds.) *Improving Inquiry in Social Science* (pp. 75-107) Hillsdale, NJ: Erlbaum.

Wiliam, D. (1996a) Meanings and consequences in standard setting, *Assessment in Education* 3(3), pp.287-307.

Wiliam, D. (1996b) Standards in examinations: a matter of trust? *The Curriculum Journal*, 7 (3), pp. 293-306.

Wiliam, D. (1998) Construct-referenced assessment of authentic tasks: alternatives to norms and criteria, Paper presented at the 24th Annual conference of the International Association for Educational Assessment – Testing and Evaluation: Confronting the Challenges of Rapid Social Change, Barbados, May 1998

Wolf, A. (2002) *Does Education Matter?* (London, Penguin Books).

Yorke, M. (1998) Assessing capability, in: J. Stephenson & M. Yorke (Eds.) *Capability and Quality in Higher Education* (London, .Kogan Page).