Studies in Higher Education *Volume 27, No. 3, 2002*

Carfax Publishing
Taylor & Francis Group

Summative Assessment in Higher Education: practices in disarray

PETER T. KNIGHT

The Open University, UK

ABSTRACT *The article begins with a view of learning and of what its assessment entails, arguing that it is helpful to distinguish between assessment systems primarily intended to provide feedout and those intended to provide feedback. Attention is then concentrated on summative, feedout, or high stakes assessment, which is supposed to be highly reliable. A number of difficulties with current practices are then identified, leading to the claim that high stakes assessment in first degrees is in such disarray that it is difficult to know what grades or classifications mean and risky to treat them as reliable. Two main lines of response are explored. The first treats this as a failure of technique, while the second adds that assessment purposes themselves have fallen into disarray, which requires a reappraisal of curriculum in addition to any technical changes that appear to be expedient.*

Learning and Assessment

Black has argued that 'researchers are faced with the difficult task of changing understanding of assessment issues, both amongst the general public and amongst policy makers' (2000, p. 407). This article argues that this is an important task for those interested in good higher education.

Although life is about learning, educational establishments are concerned with certain sorts of valued learning. The curriculum specifies the skills and understandings that are valued and, increasingly, identifies desirable self-theories (Dweck, 1999) and dispositions. So, students in higher education might be expected to understand material of importance in a subject area; to develop subject-specific and general skills; to become more confident that, with perseverance, they usually can have some success; and to be disposed to reflect and think strategically.

A great deal is known about skills learning and learning for understanding (Ohlsson, 1996), rather less about encouraging the incremental self-theories that go with persistence and strategic thinking (Dweck, 1999), and there is considerable dispute about the extent to which learning is contexted and hence transferable (Anderson *et al.*, 2000). However, there is good evidence that student achievement is related, first and foremost, to engagement (Astin, 1997). Engagement does not simply equate to the amount of involvement in and time on task, important though that is. It extends to learners' engagement in communities of practice, to their involvement in a variety of networks and to the amount and quality of interchanges with others. This is an endorsement of the neo-Piagetian position that cognitive engagements with others are powerful stimuli for learning, and of Vygotsky's analysis of learning as social acts. According to Brown & Duguid (2000), participation in communities and networks regularly sustains learning that is not easily specified in advance, cannot necessarily be measured and is often unpredictable. Important things are learned in vibrant communities that lie outwith the formal curriculum and complement it.

ISSN 0307-5079 (print)/ISSN 1470-174X (online)/02/030275-12 © 2002 Society for Research into Higher Education
DOI: 10.1080/0307507022000 0662

Notice the juxtaposition of engagement and quality. Research with schoolchildren has repeatedly found that engagement, as measured by the amount of time on task, matters, but so too do the quality of the learner's engagement (is it mindful and alert?) and the quality of the tasks (are they busy-work tasks or well-matched to the learner?) Curriculum quality also matters. Those interested in school curricula have long valued consistency, coherence and progression; what Biggs (1999) has described as an alignment between content, intended outcomes, pedagogies and assessment practices.

The award of a degree should attest that graduates have largely achieved what the curriculum specifies, with the added implication that what has been achieved is, in some measure, transferable. It would be simple if warrants to achievement could be based solely on evidence of serious engagement with curricula that were carefully designed as suites of tasks and learning processes congenial to the development of the valued dispositions, self-theories, understandings and skills. If warrants could be grounded on evidence of engagement, that would sit well with a view inspired by complexity theory (van Geert, 1994); namely that, because much about learning is indeterminate, it is best to provide the conditions for good learning and then to trust the process: generally, good learning engagements will have good outcomes. However, different conventions prevail and it is expected that warrants will be based upon judgements of achievement; on purportedly reliable assessments of student learning.

Notice that these assessments have at least three conditions to meet if they are to be consistent with the account of learning that has just been sketched.

1. They have to be faithful to the curriculum (charged with developing understandings, skills, self-theories and reflectiveness).
2. They must align with the notion that education is concerned with some degree of abstraction, generalisation or transfer.
3. They should not impede student engagement in communities of practice, but should encourage behaviours associated with good learning.

Underlying this account is the view that:

> The single, strongest influence on learning is surely the assessment procedures … even the form of an examination question or essay topics set can affect how students study … It is also important to remember that entrenched attitudes which support traditional methods of teaching and assessment are hard to change. (Entwistle, 1996, pp. 111–12)

Feedout and Feedback

When assessment certifies or warrants achievement it has a feedout function, in that the grades and classifications can then be treated as a performance indicator for the student, department, institution, employer, funding body, quality agency or compilers of league tables. So important are those feedout functions that such assessment is often called high stakes or summative assessment, and great emphasis is consequently put on making it robust. Reliability is a crucial feature of robust assessment. Careless or capricious feedout is unethical and can be challenged. Assessment can also have a feedback function when it is intended to evoke information to help further learning. The argument that this is a very important function of assessment is not to be developed here, because its full significance cannot be disclosed if it is assumed that all is well with summative assessment practices. If it can be argued that summative assessment is in disarray, then the feedback functions can be reappraised, thereby putting consideration of the place of assessment in (higher) education in a fresh light. This is not the place for that project. This article is concerned with the more modest task of suggesting that summative assessment practices are in disarray.

While all assessments involve questions about validity, reliability, usefulness and cost, different assessment intentions lead to such different answers that there is a strong case for regarding feedout and feedback assessment as different systems. When the intention is to create feedback, learners need to be willing participants who disclose their uncertainties, errors and lacunae. Where feedout is the goal, disclosure is displaced by deception. Players in assessment games need to know which rules apply and there is little room to apply both sets to the same task. Not only do summative and formative assessments instantiate different rules of engagement, because they have different intentions, they also have different rules of evidence. Similarities are identified in the following list, but the view that there are profound differences between feedout and feedback shapes the analysis of summative assessment that shortly follows.

1. Assessors look for evidence of achievement. Evidence is a teleological concept, which means that it involves identifying some data as relevant to specified goals and the criteria that derive from them. In higher education, the curriculum specifies them. The goals or criteria may be simple or complex and expressed more or less precisely. This means that there is a range of certainty about what counts as good—or fair—evidence. With fuzzy, complex criteria there is considerable scope for disagreement about what counts as evidence of attainment.
2. Judgements are made about the match between evidence and criteria. With simple goals and very precise criteria no judgement is really needed—these are no-inference or low-inference cases when highly reliable conclusions follow. In many cases conclusions are much less reliable, because the goodness of fit between evidence and criteria has to be inferred. Attempts are often made to increase reliability by reducing the inferential load through establishing inferential rules, typically in the form of achievement criteria. This is not a complete solution, though, because no statement of a rule (including a rule stating a criterion) lays down (entails) its meaning, interpretation and application (Kripke, 1982). This is all the more so when the goal in question—critical thinking, for example—is itself polymorphous.
3. Judgement invokes communication. Some assessment judgements are relatively low stakes, may be fuzzy and exploratory and conversational in character, as in a discussion amongst peers or in dialogue with a tutor. In such cases there is often space on all sides for meanings to be negotiated and clarified. High stakes assessments are marked by unequal power relations between assessor and assessed, the format is not conversational and the judgement may be in a highly symbolic form, such as a number or letter grade, that is capable of multiple decodings. Understood as communications, summative and formative assessment are very different. The former is about conveying information and might be understood with reference to the theories of Shannon and Weaver (Fiske, 1990), while the latter is more hermeneutic and could be illuminated by the work of theorists such as Gadamer (Moran, 2000) or Luhmann (1995). Some implications for summative assessment are pursued later in this article.
4. Judgements are economic processes. They have direct money costs (as when people are paid a fee for assessment services), indirect costs (as when academic staff manage the extra work of criteria-referenced grading by working longer), and opportunity costs, which may be the most significant. When people are assessing or being assessed, it is at the expense of doing other things. Bearing in mind that the costs include preparing for assessment and

> recording the outcomes, as well as the smaller costs of actually assessing/being assessed, it becomes evident that even run-of-the-mill assessments can be expensive. Stress (amongst teachers and learners) and dysfunctional behaviours (of which surface learning strategies can be one) are common consequences. And although formative assessment is intended to enrich the learning economy, summative assessment's contribution to learning is less palpable.

The argument now concentrates on summative assessment, suggesting two things: first, that it is being stretched to cover learning goals that resist robust, reliable and affordable summation; and secondly, that these assessment practices are shot through with contradictions such that the data they provide signify less than people tend to assume.

Summative Assessment in Disarray

The complex history of psychometrics shows that summative assessment has always been a vexed business, so what are the grounds for claiming that high stakes assessment is *now* in disarray? There are several complementary answers. One is that higher education institutions are generally expected to have learning goals that are far more extensive and complex than mastery of subject matter alone, *and* that they are being held to account for student achievement in terms of those goals. Thus, in the UK, they are now accountable for student employability. This invites the extension of high stakes assessments to learning achievements, to which consensus about high stakes assessment no longer applies. Secondly, at the same time, a greater range of assessment techniques has come into currency, which has introduced substantial practical and theoretical problems, with the comparability and aggregation of performances judged by different assessment methods. Thirdly, public sector services are nowadays marked by low-trust management systems, when once there would have been a greater readiness to trust that good people engaged on worthwhile activities would learn the sorts of things that were intended. Assessment is supposed to supply evidence to bridge the trust gap. Fourthly, the eternal concern with value for money has taken a rationalist turn, with the belief that it is prudent to specify objectives, measure inputs, assess performance in terms of those objectives, allocate the next round of resources to efficient providers and apply sanctions to the less efficient. This approach is also assessment-intensive, as is a related concern with the maintenance of standards and the enhancement of quality. All of these rationalist, low-trust, risk-averse approaches feed on assessment data, and, when it becomes evident that summative assessment systems cannot provide the robust performance indicators upon which they depend, then control, accountability and legitimacy are all compromised.

This pressure to extend the feedout function of assessment compromises the three conditions good assessment must meet (see above). Reliability, which must suffuse the process of warranting achievement, is greatest when there is uncontentious evidence of achievement on many occasions that comes from low-inference assessment routines. This quest for reliability tends to skew assessment towards the assessment of simple and unambiguous achievements, and considerations of cost add to the skew away from judgements of complex learning. To put it another way, high stakes assessments have trouble with the complex ambitions of higher education curricula and may actually impede them (Boud, 1995). To compound matters, it should be understood that summative assessment may not be able to deliver what it is widely supposed to. Notes on 10 areas of difficulty follow, implying that the costs of summative assessment are not always offset by what it is able to deliver.

Knowledge and Knowing

Assessment involves making assumptions about what exists, what it is like and how we might know about it. For example, if skills are nothing more than convenient terms for social practices that are decidedly situation-specific, and hence changeable (Holmes, 2001), then it will be frustrating to try to assess skills as if they were real, generalisable achievements. Again, what some take to be a psychological property, such as self-esteem, that is measurable and has explanatory powers may, in fact, be no more than a non-stable self-evaluation with no explanatory powers (Harré, 1998). The two views have different implications for what we might wish to assess and our chances of success. The importance of establishing what might be possible can be further illustrated by considering the use of criteria to make assessment practices more reliable. Although it may be sound common sense to have assessment criteria, it has been seen that there are philosophical and psychological objections to the claim that criteria can pre-specify the outcomes of good learning (Kripke, 1982). Furthermore, there are objections to the claim that we have conscious access to all that we know, which raises obvious problems with trying to assess it. There is a strong line of thought which holds that much of what we know is tacit and distributed, and that much learning is non-formal. Eraut (2000) gives a fine account of the difficulties that attach to all claims to describe and, by implication, to assess, tacit knowing, the significance being that much of the tacit and distributed knowing derived from non-formal learning may elude capture for summative assessment purposes. Since there is agreement that these sorts of knowings can be very important, this is an important limitation.

The Limits of Reliability

Plainly, fictional objects of assessment cannot be assessed with validity, and where validity is lacking, reliability is compromised. So, were skills to be fictions, there would be interesting validity and reliability issues attaching to all efforts to assess them. It has already been suggested that some of the other qualities that higher education institutions might claim to promote (self-motivation, for example) cannot be assessed reliably (and affordably), and that some cannot be ethically assessed either. Besides, the assessment of complex or divergent achievements is inherently unreliable. Essays are probably the most familiar way of assessing complex cognitive achievements, but, familiarity notwithstanding, Fleming (1999) identifies some of the many sources of unreliability in essay marking, while Breland (1999) shows the problems of reducing unreliability when simple pieces of student writing are involved. More generally, Heywood (2000) notes the unreliability of many assessment routines and compares higher education practices unfavourably with those used in A level examinations in England.

The Stability of Assessment Judgements

Suppose that the first two sets of objections did not apply in a particular case, such as the ability to work safely in laboratory conditions. One might make an inference about it on the basis of one observation, but it would be unwise to generalise from it because it is a very small sample of all the possible occasions in which the subject might work in a laboratory. Repeated observations are necessary before claiming that the observed behaviour is likely to be stable. This is important because if a higher education institution wishes to warrant achievement, then the warrant should be based on several assessors judging different instances of it. That can hardly be done in a single module, but it might be done if there was an assessment plan covering a complete degree programme, so that the important judgement that someone works safely in a laboratory could be based on three or four years of targeted assessments. However, programmes have widely been deconstructed by modularisation and increased student choice, which makes desirable summative assessment practice somewhat elusive.

The Transferability of Achievement

The achievements that grades or degree classes signify may not be very transferable: Many psychologists hold transfer to be an achievement in its own right, not something that flows freely and easily, except in familiar settings where specific transfer heuristics have been routinised (Anderson *et al.*, 2000). The degree to which what is learned may be transferable has a lot to do with how it is learned, with whether metacognition has been instantiated to foster a propensity to transfer. That depends very much on the learning processes, about which grades and degree classes are usually silent. Nor do scores and grades say anything about the learner's ability to perform independently or in novel contexts. They may indicate a performance achieved with the help of plenty of scaffolding or with none. Even where two products are fairly awarded the same grade, there is a real difference between an achievement where the task has been well defined, procedures for success have been laid down and plentiful guidance has been available, and another where no scaffolding has been provided. In other words, it is proper to doubt any assumption that warrants achievements that the learner can readily and independently transfer to fresh settings.

Limitations to Criteria-referencing

Although criteria-referenced assessment has many strengths, particularly compared to norm-referencing, it is important to insist that benchmarks, specifications, criteria and learning outcomes do not and cannot make summative assessment reliable, may limit its validity and certainly compound its costs. For example, educational criteria are necessarily imprecise unless they refer to highly determined, even trivial achievements, as with the lower levels of the English National Vocational Qualifications. Trainers may be able to develop and use precise-looking criteria, but educators work with fuzzy learning outcomes. Even 'precise' criteria are fuzzy to the extent: (i) that their meanings emerge in local communities of practice; and (ii) in the context of specific tasks (Wolf, 1997). It is hardly surprising, then, that difficulties are reported in getting agreement on criteria and their application in a subject (Greatorex, 1999) and in a school (Price & Rust, 1999). Notwithstanding attempts to harmonise grading criteria, there remain significant variations between groups of higher education institutions and between subject communities. Although criteria-referenced grading may be good for student learning and equity in a community of practice, differences in the criteria used prevent, even impede, communication *between* communities; and make it impossible to be sure what any warrant means, since it is not possible to know what criteria have been used, what meanings have attached to them and how they have been used.

Assessment and Curriculum Skew

This point has been briefly made but is important enough to bear elaboration. High stakes assessments, of the sorts that appear on transcripts and that lead to awards, have to be robust enough to stand up to legal challenge, so they tend to rest on assessments of things that people (often wrongly) believe can be judged reliably. This distorts the curriculum in two ways. First, what is subject to high stakes assessment gets serious attention and the rest does not. Secondly, achievements that are not warranted by high stakes assessment are neither recorded nor celebrated. The enacted curriculum becomes what high stakes judgements cover. Non-authentic assessments produce non-authentic curriculum, regardless of what the validated curriculum claims. This is serious in systems mandated to evoke complex student learning, as when governments expect universities to develop four dimensions of employability in students; namely, subject understanding, skills, robust self-theories and reflective or metacognitive casts of mind. There is a real danger that the frustrations of trying to assess such accomplishments in reliable ways will lead to the use of national, content-free tests, such as American College Test (ACT), Graduate Record Examination (GRE) and Graduate Management Admissions Test (GMAT), as proxies for sound authentic assessment. Not only is their predictive validity in doubt (Sternberg, 1997), but if students concentrate on becoming test-smart, the tests' consequential validity decreases because they actually distract students from the curriculum designed to teach those things that the tests claim to measure by proxy.

The Misuse of Number

It is seldom appreciated that although feedout assessment data are usually presented numerically, they should not routinely be treated 'numerically'. True numerical data have well-defined properties that most data on human behaviour do not have (Mitchell, 1997). At the very least, this means that conclusions based on statistical routines intended for truly numerical data may not be valid. In other words, beware of the numbers created by summative assessment and mistrust conclusions based on the transformation or manipulation of those numbers.

To this general point can be added some specific problems with numerical data. In the UK, some subjects use grading scales, others percentages. Scales are not necessarily commensurable and percentage marks have different meanings in different subjects (Yorke *et al.*, 2000). Where awards are based on mean marks, students in subjects using the full 0–100 mark range will not be treated the same as those in subjects using a more restricted range—30–75 is common in humanities and social sciences. There are similar variations in the ways that scales are used. In other words, different rules operate for the allocation of numbers to achievement, which means that: (i) there are difficulties in trying to combine numbers produced by different conventions; and (ii) it is not always clear which conventions lie behind any given set of numbers. Furthermore, student achievement is usually reduced to a six-point scale for degree classification purposes (first ... fail), and most marks fall into two categories (2:1 and 2:2). This limited range constrains value-added calculations, which are further compromised by the kurtosis coming from the increase in the proportion of 2:1 grades.

The Opacity of Number

Some grades or classifications are based only on examinations, some only on coursework, and some on varying mixes of the two. Different weightings can produce different grades and classifications (Dalziel, 1998), with evidence emerging that students tend to score more highly on coursework assessments (Yorke *et al.*, 2000). This means that the same grade might describe skill in examinations *or* perseverance in coursework *or* a blend of both, and that a different grade might have been awarded on the basis of the same set of marks combined using a different formula. In fact, there is considerable variation in the amount of work that is assessed for grading or classification (Yorke, 1999), which is significant, because students who feel overworked may adopt 'surface' learning approaches, which should not be compatible, in humanities and social sciences, with the best marks (Chambers, 1992). The implication is that moderate marks may represent overload or moderate effort/aptitude. There is also some inconsistency between groups of higher education institutions in the ways that scores from different years of study are weighted. It is not clear whether a degree classification describes students' sustained performance across the programme, the level they reached at the end of it, or some unknown blend of the two.

The main point here is that grades, scores and degree classes are uninformative. Degree transcripts could be an answer, but one North American study found that many did not make it clear what students had learned (Adelman, 1990). More informative transcripts can sprawl into long lists of statements of achievement, and be ignored because they are unwieldy.

Process-blindness

Scores and grades are silent about the learning processes involved. This is important on at least two grounds. One is that anyone wishing to judge the robustness of an achievement needs to know something about its circumstances. If you tell me that someone has repeatedly shown that they can solve problems, and I find that problem-solving has been taught and learned as the manipulation of numbers according to learned algorithms, I may be less impressed than if I hear that it has been developed through engagement with a series of 'fuzzy', authentic tasks. The quality of the learning processes—of a programme's process standards—tells something about the robustness of what has been learned.

The implication is that formally equivalent qualifications may not attest to equivalent learning. The second point extends it by referring to Brown & Duguid's (2000) claim that the processes involved in getting a degree are important because much learning comes with the quality of interactions in the communities to which students belong. That is to say that learning processes affect the quality of what is learned *and* what is learned as well.

Utility

Summative assessments that appear to speak reliably about some achievements at given points in the undergraduate years can be moderate or poor predictors of career achievement (Sternberg, 1997). This poor predictive validity rather diminishes the value of these assessments as feedout. Employers, who might be expected to rely on summative assessment data, perhaps recognise this. Some appear dismissive of awards from low-status higher education institutions, and others appear to appoint on the basis of evidence other than degree results. If employers, who may be acting rationally in the light of their situations and problems with summative assessment data, mistrust assessment data, questions arise about their exchange and use values.

Responses

One response to such a litany is to look at ways of resolving some difficulties and living with others. There are achievements that can be fairly well defined and assessed with good levels of reliability at reasonable cost across different departments teaching the same subject in their own ways. There is certainly no shortage of ideas about ways of improving summative assessments. For example, existing procedures could be tightened up by a more diligent pursuit of plagiarism and malpractice, or by ensuring that all work is double-marked by graders who could not know the identity, sex or race of the candidate. Holroyd (2000, p. 42) suggests, amongst other things, providing continuing training for academic staff in assessment and its development; 'effective interaction with all the assessment stakeholders'; a code of practice; and 'a commitment to critical reflection ... and research on assessment'. North American work, such as Walvoord & Anderson's (1998) *Effective Grading*, offers a lot more detailed advice.

Yet, too much may be expected of steps like these. Take, for example, the matter of improving assessment training. Ideally, no one would prefer untrained assessors to trained ones, but that still leaves questions open about what the training should involve and what it might expect to achieve. Consider standardisation meetings in England for A level examiners in, say, history, which can be regarded as specialised assessment training events for teachers who will already have considerable experience of marking students' complex essays. The Chief Examiner will already have marked a batch of scripts, explored an emerging marking scheme with senior colleagues and refined it in the light of their experiences of trying to use it. The assistant examiners then meet for a day to master and refine it. They are briefed on it, questions are discussed and then they practise using it, answer by answer. Provisional marks are collected, the scheme is revised and examiners who are marking too high or too low are advised on how to get into line. During the day a body of case law develops and enters into the marking process. Assistant examiners then go away and begin marking. Samples are moderated and examiners whose marking is out of line are removed. This is a system for grading complex work that is about as reliable as it is possible to get at an affordable price. Even supposing that assessor training in higher education institutions were this purposeful and rigorous (and there are many reasons, not least ones to do with costs and the priority of research, to doubt it), could technical reforms of this sort settle the issues outlined in the previous section?

The argument here is that they would not, and the claim is developed with reference to some theories of communication. It is a curt summary of a complex field, and readers wishing to know more might turn to an introductory text such as Fiske (1990). Some communication theories raise questions about the goodness of fit between the message sent and that received. They direct attention to ways of improving the strength and clarity of the signal, and of reducing ambiguities and noise. Applying this perspective helps us to see technical reforms as attempts to clarify and amplify the signal (information about students' real achievements) and minimise noise, which takes the form of ambiguous or erroneous information. This assumes, of course, that background and noise can be filtered out so that pure information about real achievement can be transmitted. However, I have said that there are philosophers and psychologists who have serious reservations about this. Modern philosophies that treat facticity and historicity as the essence of being, not as noise obscuring transcendental essences, are complemented by psychologies of situated cognition that ask how far any general faculties that might exist, such as *g* (general intelligence), could be freed from specific contexts and contents. These views call into question the idea that signal and noise are separable. Other theories of communication see the 'receiver' creating meaning out of the information that is received. Meanings are not received in the way that the postal service delivers a package. A sign (a grade, for example) is in some relation to an object (a performance or achievement), and relates to an interpretant (a meaning: for example, 'this person isn't very clever'). To repeat, these semiotic theories have the receivers as active co-creators of the interpretant—of meaning. This implies that higher education might make technical reforms to try to improve the quality of a restricted range of feedout without stakeholders creating meanings of the sorts that they want or that higher education intended.

Furthermore, the more diversity there is in the sending systems, the greater the likely range of interpretants and uncertainty about them. So, where an assessment system is long-established and based on a common curriculum, as is the case with A levels in England, then the range of interpretants about, say, a C grade is likely to be limited and 'aberrant decodings' exceptional. On the other hand, in the USA, the interpretants of high school grades from an out-of-state school are less constrained by convention. Given that higher education does not have a common curriculum, and that even within disciplines consensus can be elusive, it follows that, regardless of anything done to improve signal quality, there are problems with the meanings that might be attributed to grades and degree classifications, because powerful conventions do not operate as they do with A levels and other trusted public examinations. The massification and diversification of higher education systems, allied to the turbulence and differentiation of modern times, have disrupted the old conventions by which meaning was ascribed to assessment symbols. This is compounded by the spread of the belief that higher education institutions should promote complex achievements—skills, self-theories and metacognitive acuity—which are especially hard to reconcile with a wish to secure cheap and reliable judgements, because these achievements tend to be indeterminate and ambiguous. Enthusiasm for validity in the shape of the assessment of authentic achievement (Cuming & Maxwell, 1999) exacerbates matters. The suggestion is that higher education might do a lot to try to clarify the signals it intends to send, but problems of interpretation may lie not so much with the signs as with the 'receivers'.

In the *Nicomachean Ethics*, Aristotle advises us not to expect more precision than the subject admits of, which is a good precept to apply to summative assessment. In many ways it cannot deliver the precision and certainty that managerialist discourses and common sense expect. Technical reforms and plenty of resources might help us to get more precise, reliable, robust and informative feedout, but it has been argued that some summative assessment problems are problems only in the sense that things are being expected of summative assessment that it cannot do or, in some cases, cannot do at a price that higher education institutions can pay. This is contrary to the position implicit in most 1990s writings on assessment, which treat assessment problems as outcomes of methodological ignorance. The position here is that discussion of methods and their application is sterile without deliberation on (or deconstruction of) assessment-in-curriculum; which is to say that radical thinking is needed about what summative assessment is for, who it is for, what it can do, what it cannot do cheaply and what it ought not to be asked to do at all. Three lines of thought that are attracting me are as follows.

1. If higher education restricted feedout to achievements about which affordable, reliable and fair judgements could be made, then this narrowing of the range would be a basis from which to work at encouraging shared understandings, and be a release from expensive attempts to do the impossible. The argument is that this is a necessary backward step.
2. There is a need for systems of formative assessment that engage students with feedback about their work in order to signal what else is valued in the curriculum, what might count as fair evidence of achievement in those terms, and to indicate directions for further learning. The aim of formative assessment would be to stimulate discourse characterised by listening and exchange, with as little imbalance of power as possible. The inspiration is Gadamer's remark that 'a genuine conversation is never the one we wanted to conduct' (cited in Moran, 2000; p. 249). This has considerable implications for curriculum planning.
3. Higher education institutions may be ill advised to provide feedout about many achievements that are of interest to employers and other stakeholders, but that does not mean that they ought not to help those stakeholders to make judgements about students' achievements in such areas. Since formative assessment would help students to recognise

> their achievements, they would be in a position to make their own claims to things that higher education institutions did not warrant. Employers could satisfy themselves about claims by asking for evidence from the portfolios that students would develop as adjuncts to a programme-wide approach to learning and formative assessment. If higher education institutions also provided summaries of the key learning, teaching and assessment processes associated with each programme of studies—of their process standards—stakeholders would be more able to scrutinise claims and warrants alike. Those wanting to know about higher education institutions contributions to student achievements would use connoisseurship to appraise institutional claims to add value. Mounting evidence about the shortcomings of attempts to measure the added value that comes from high schools should make us cautious about using statistical techniques to judge departmental, faculty or institutional effectiveness.

I have argued that summative assessment is in disarray, said that formative assessment needs attention and implied that so too does the nature of curriculum in higher education. I have claimed that there is little to be had from discourses that blame problems on defective methods, less-than-competent teachers, scant resources, or failures to apply the linear systems thinking of rational management practice. The deficiencies of which they speak have long been with us. Better, I suggest, to explore assessment as complex systems of communication, as practices of sense-making and claim-making. This is about placing psychometrics under erasure while revaluing assessment practices as primarily communicative practices.

Acknowledgements

This work was done at Lancaster University. I benefited considerably from comments from Kathryn Ecclestone (University of Newcastle), Mantz Yorke (Liverpool John Moores University) and the anonymous referee who engaged vigorously with an earlier version of the paper. I suspect that each would still wish this to read differently.

Correspondence: Peter T. Knight, COBE, The Open University, Milton Keynes, Buckinghamshire MK7 6AA, UK.

REFERENCES

ADELMAN, C. (Ed.) (1990) *A College Course Map: taxonomy and transcript data* (Washington, DC, US Government Printing Office).

ANDERSON, J.R., GREENE, J.G., REDER, L.M. & SIMON, H.A. (2000) Perspective on learning thinking and activity, *Educational Researcher*, 29(4), pp. 11–13.

ASTIN, A.W. (1997) *Four Years that Matter: the college experience twenty years on* (San Francisco, CA, Jossey–Bass).

BIGGS, J. (1999) *Teaching for Quality Learning at University* (Buckingham, Society for Research into Higher Education and the Open University Press).

BLACK, P. (2000) Research and the development of educational assessment, *Oxford Review of Education*, 26, pp. 407–419.

BOUD, D. (1995) Assessment and learning: contradictory or complementary?, in P.T. KNIGHT (Ed.) *Assessment for Learning in Higher Education* pp. 35–48 (London, Kogan Page).

BRELAND, H.M. (1999) From 2 to 3Rs: the expanding use of writing in admissions, in: S.J. MESSICK (Ed.) (1999) *Assessment in Higher Education: issues of access, quality, student development and public policy*, pp. 91–111 (Mahwah, NJ, Lawrence Erlbaum Associates).

BROWN, J.S. & DUGUID, P. (2000) *The Social Life of Information*, (Cambridge, MA, Harvard University Press).

CHAMBERS, E. (1992) Work load and the quality of student learning, *Studies in Higher Education*, 17, pp. 141–153.

CUMING, J. & MAXWELL, G. (1999) Contextualising authentic assessment, *Assessment in Higher Education*, 6, pp. 177–194.

DALZIEL, J. (1998) Using marks to assess student performance, *Assessment and Evaluation in Higher Education*, 23, pp. 351–366.

DWECK, C. (1999) *Self-theories: their role in motivation, personality and development* (Philadelphia, PA, Psychology Press).

ENTWISTLE, N. (1996) Recent research on student learning, in: J. TAIT & P. KNIGHT (Eds) *The Management of Independent Learning*, pp. 97–112 (London, Kogan Page).

FRAUT, M. (2000) Non-formal learning and tacit knowledge in professional work, *British Journal of Educational Psychology*, 70, pp. 113–136.

FISKE, J. (1990) *Introduction to Communication Studies*, 2nd edn. (London, Routledge).

FLEMING, N. (1999) Biases in marking students' written work, in: S. BROWN & A. GLASNER (Eds) *Assessment Matters in Higher Education*, pp. 83–92 (Buckingham, Society for Research into Higher Education & Open University Press).

GREATOREX, J. (1999) Generic descriptors: a health check, *Quality in Higher Education*, 5, pp. 155–166.

HARRE, R. (1998) *The Singular Self* (London, Sage).

HEYWOOD, J. (2000) *Assessment in Higher Education*, (London, Jessica Kingsley).

HOLMES, L. (2001) Reconsidering graduate employability: the graduate identity approach, *Quality in Higher Education*, 7, pp. 111–120.

HOLROYD, C. (2000) Are assessors professional? *Active Learning in Higher Education*, 1, pp. 28–44.

KRIPKE, S.A. (1982) *Wittgenstein on Rules and Private Language* (Oxford, Blackwell).

LUHMANN, N. (1995) *Social Systems*, tr. J. BEDNARZ (Standford, CA, Standford University Press).

MITCHELL, J. (1997) Quantitative science and the definition of measurement in psychology, *British Journal of Psychology*, 88, pp. 355–383.

MORAN, D. (2000) *Introduction to Phenomenology* (London, Routledge).

OHLSSON, S. (1996) Learning to do and learning to understand, in: P. REIMENN & H. SPADA (Eds) *Learning in Humans and Machines: towards an interdisciplinary learning science*, pp. 37–62 (London, Pergamon).

PRICE, M. & RUST, C. (1999) The experience of introducing a common criteria assessment grid across an academic department, *Quality in Higher Education*, 5 pp. 133–144.

STERNBERG, R.J. (1997) *Successful Intelligence* (New York, Plume).

VAN GEERT, P. (1994) *Dynamic Systems of Development: change between complexity and chaos* (Hemel Hempstead, Harvester Wheatsheaf).

WALVOORD, B.E. & ANDERSON, V.J. (1998) *Effective Grading: a tool for learning and assessment* (San Francisco, CA, Jossey–Bass).

WOLF, A. (1997) *Assessment in higher education and the role of 'Graduateness'* (London, Higher Education Quality Council).

YORKE, M. (1999) Benchmarking academic standards in the UK, *Tertiary Education and Management*, 5, pp. 81–96.

YORKE, M., BRIDGES, P. & WOOLF, H. (2000) Mark distributions and marking practices in UK higher education, *Active Learning in Higher Education*, 1, pp. 7–27.