The Sixth *QHE* Seminar

The End of Quality?

Birmingham, 25–26 May, 2001

The Achilles' heel of quality: the assessment of student learning

Peter T. Knight,

Educational Research, Lancaster University

Theme 1

Abstract

This paper explores the dependability of assessments of student achievement when used as performance indicators for internal and external quality monitoring (IQM and EQM). Problems are identified that jeopardise attempts to monitor, control and enhance quality in higher education. Responses are suggested with preference being given to a radical approach based on accepting that reliable national data about complex student achievements are not to be had. It is argued that this means that reliance on EQM is unwise and that more attention should be paid to internal quality enhancement.

Introduction

> As someone who has spent his entire career doing research writing and thinking about educational testing and assessment issues, I would like to conclude by summarizing a compelling case showing that the major uses of tests for student and school accountability over the past 50 years have improved education and student learning in dramatic ways. Unfortunately, that is not my conclusion. (Linn, 2000, 14)

If we ask what is the end of quality, what purposes are served by the global concern with the quality of higher education , then problems come before answers. One is that there is a lack of agreement about the meaning of quality (we would expect nothing less of academics). Harvey (Harvey and Green, 1995) has identified five

forms of quality and Knight and Trowler (2000) have described two ends of a spectrum of meanings (Table 1). Different discourses sustain different practices and imply different answers to questions about the end of quality (the purpose) and the end of quality (the sustainability of quality enhancement measures).

Type I approaches to quality assume the existence of good (that is to say, reliable and valid) performance indicators (PIs), as the next section demonstrates.

Table 1. Two types of quality

Type I	Type II
Efficiency: 'lean and mean'	Effectiveness: spaces and redundancy
Emphasis on measurables, typically outcomes	Emphasis on processes
Binding, well-specified procedures	Binding, well-rehearsed goals: procedures open
Tight coupling: hierarchies and low-trust cultures	Loose coupling: workgroups and high trust cultures
Compliance: errors punished	Creativity: errors are necessary for learning
Motivate by rewards and punishments: extrinsic	Self-actualisation and fulfilment matter: intrinsic motivation
In practice, emphasis on single-loop learning	Signs of double-loop learning (and more besides?)
Linear view of the social world: direct cause-effect connection	Complexity view: attractors constrain causes, interactions make effects unpredictable
'Rational' view of communication and planning	Communication as sensemaking, planning is far from rational

Assessment of students

Assessment in quality monitoring

Some of the reasons why summative assessment data of student performance are not trustworthy performance indicators are summarised in Table 2. They are unreliable and routinely mismanipulated, incomplete and generally uninformative representations of student achievements. It is imprudent to rank courses and institutions on the basis of them, which compromises most Type I IQM and EQM systems.

Table 2. Some reasons why higher education assessment data are poor performance indicators

Assessment problem	Why it matters
A. The misleading power of numbers	
1. Some academic subjects use grading scales, others per centages. Scales are not necessarily commensurable and practices in the use of percentages are not consistent (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.
2. In the UK, student achievement is shown on a six point scale (1st ... fail) and most marks fall into two categories (2:1 and 2:2).	This limited range constrains value-added calculations and the increasingly skewed distribution caused by the increase in the proportion of 2:1 grades awarded compounds it.
3. There is no agreement about how to calculate added-value scores (Saunders, 1999).	If assessment data are used as PIs, it is only fair to ignore raw scores and make added-value calculations. The lack of an agreed method compromises the attempt to use assessment data as PIs.
4. Although student assessment data are usually presented numerically, they should not be treated 'numerically' (that is, as interval or ratio scale data).	Assessments of human achievement are likely to produce ordinal data which should not be analysed as interval or ratio scale data. Conclusions based on routines intended for 'higher-order' numerical data may not be valid (Cliff, 1996; Mitchell, 1997)
B. The endemic unreliability of assessments.	
5. The assessment of complex or divergent achievements is inherently unreliable, although reliability can often be increased, to some degree, if enough costly effort is put in by using well-trained multiple markers, making repeated observations, etc..	Fleming (1999) identifies some of the sources of unreliability in essay marking and Breland (1999) considers ways of reducing it. Heywood (2000) notes the unreliability of many assessment routines and remarks on the expense of enhancing reliability, where it can, indeed, be enhanced. Higher education institutions (HEIs) do not usually achieve the levels of reliability of A-level examinations.
6. Some of the qualities that HEIs might claim to promote (self-motivation, for example) cannot be assessed reliably and affordably. Some cannot be ethically assessed.	Governments and employers increasingly want HEIs to deliver highly employable graduates and often define employability partly in terms of the possession of individual qualities, attitudes and beliefs. If these cannot be reliably assessed then alternative ways are needed to communicate students' complex achievements to stakeholders in higher education.
C. Summative assessment data as good descriptors of achievement.	
7. 'High stakes' assessments, of the sorts that appear on transcripts and that lead to awards, have to be robust enough stand up to legal challenge. That means that awards and transcripts tend to rest on assessments of things that can be judged reliably.	High stakes assessments skew 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.
8. Test scores and highly reliable assessments tend to give reliable data about simple achievements that are of secondary importance.	National, content-free tests (ACT, GRE and GMAT, which are in use in North America) rely on self-contained items that have correct answers. They are not assessments of authentic achievements and may not be good predictors of performance in complex settings (Cuming & Maxwell, 1999).
9. It is not usually clear how stable are the achievements to which grades or classification attest.	Repeated observations of an achievement are necessary before claiming that it is likely to be stable.

10. The achievements that grades or classes signify may not be very transferable.	To some extent, achievements are contexted (Anderson *et al*., 2000). The degree to which they may be transferable has a lot to do with the learning processes, about which grades and classes are usually silent.
11. Scores and grades do not indicate the extent to which achievements are autonomous	A 'grade-point average' may indicate a performance achieved with the help of plenty of scaffolding or with none. Notwithstanding the grade fairly awarded to the product, the achievements are very different.
12. Scores and grades are silent about the learning processes involved.	Brown and Duguid (2000) say that the processes involved in getting a degree are important. The quality of interactions in the communities to which students belong matters because much learning happens by participating in them.
D. The comparability of assessments	
13. 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). Students tend score more highly on coursework assessments (Yorke, *et al*., 2000).	This implies that module grades and degree classes are not comparable. A grade might describe skill in examinations or perseverance in coursework but a different grade might have been awarded if the same marks had been earned but combined using a different formula.
14. There is some inconsistency between groups of HEIs 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, or only the level they reached at the end of it.
15. In Britain there are different rules for determining class of degree.	Some HEIs base awards on mean marks. Others count the number of modules passed at specified levels. Rules for discounting low outlier marks also vary.
16. There is considerable variation in the amount of work that is assessed for grading or classification (Yorke, 1999; Knight, 2000).	Students who feel that they are overworked may adopt 'surface' learning approaches, which should not be compatible, in humanities and social sciences, with the best marks (Chambers, 1992). Moderate marks may represent overload or moderate effort/aptitude.
17. The grading criteria used in one HEI and/or subject tend to be distinctive.	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.
18. Difficulties are reported in getting agreement on criteria and their application in a subject (Greatorex, 1999;Woolf *et al*., 1999) and in a School (Price & Rust, 1999).	It has sometimes been assumed that the previous difficulty can be resolved by settling for agreement at the level of the community of practice (department, School or discipline). However, it is still difficult to forge agreement about the appropriateness, meaning, and application of criteria.
19. Educational criteria are necessarily imprecise unless they refer to highly determined, even trivial achievements, as with lower-level NVQs.	Trainers may be able to develop and use precise-looking criteria. Educators work with fuzzy learning outcomes. Even 'precise' criteria are fuzzy to the extent (i) that their meanings emerge in local communities of practice (ii) in the context of specific tasks (Wolf, 1997).
20. There are substantial philosophical and psychological objections to the claim that criteria can pre-specify the outcomes of good learning.	The more that criteria are used to identify what is to be valued, the more they constrain learning processes and outcomes to convergent paths. Powerful though this learning may be, there is a view that higher education should be about more besides, even (especially) if employability is a goal.

E. The usability of assessment data	
21. Even North American degree transcripts do not make it clear what students have learned (Adelman, 1990).	Grades, scores and degree classes contain very little information indeed.
22. Even if fairly reliable data can be gathered about a good range of student achievements, it is hard to report.	Fairly reliable assessments of complex learning tend to rely on repeated criteria-related judgements. These can either be summed up by numbers (in which case the explanations of what the numbers mean can get ignored) or expressed as a long list of statements of achievement (which may be ignored because they are unwieldy).
23. Employers may screen job applicants on the basis of degree class but appoint on the basis of other evidence. Some appear dismissive of awards from low-status HEIs.	If employers appear to mistrust assessment data, whether in the form of degree classification or grade transcript, that raises questions about their suitability as performance indicators.

Towards reliable assessment?

There is no doubt that some of these difficulties can be eased by simple investments in better training, clearer criteria, the use of more assessors and by more assessments but my claim is that the task is much more complex than it may appear. Three theories are used to reappraise the problems and ways of reducing them.

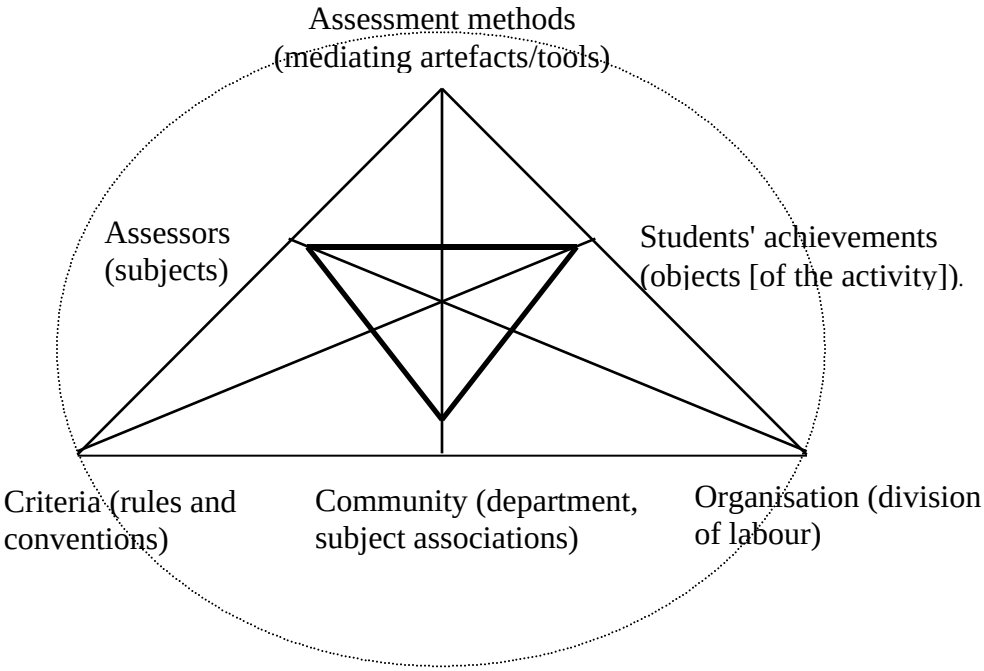

Figure 1: The assessment of achievement as an activity system

The first is ‘activity system’ theory (Engestrom, 1990). Figure 1 is based on Engestrom’s sketch but changes his notation (shown in parentheses) to illuminate the case of assessment. An activity, such as assessing students’ achievements, is seen as

the interplay of two sets of three elements, subjects, objects and community; rules, mediating artefacts and division of labour. A succinct description of an activity system is that it comprises

> ...a number of basic elements, including a given practitioner or *subject*, the *object* or motive of the activity, its *mediating artefacts* (e.g., tools, signs and symbols), the *rules* generally followed in carrying out the activity, the *community* of co-workers and colleagues involved in the activity, and the *division of labour* within the activity. (Hart-Landsberg *et al*., 1992, p. 7)

The activity of assessment is, then, understood as a system comprising six elements, each of which can be more or less complex. For example, if the activity in question is the assessment of information retained by learners, then the activity system can be a tightly-coupled orchestration of six unproblematic elements. There is little room for uncertainty about the methods, criteria or achievement and no reason why the six elements should not operate harmoniously. Well-established conventions about the assessment of information favour the generation of reliable data about students' information retention.

The assessment of problem-working is very different because there are legitimate differences of opinion about how each element should be appropriately configured (what criteria to use? who should assess? how best to organise it?). Elements cease to be points of consensus and become sites of disputed discourses. There is also little consensus about the assessment of knowing, which increases the uncertainty in the system. As a result, the six complex elements are loosely-coupled at best. At worst, they compete with each other.

The two triangles in Figure 1 are surrounded by a broken circle to indicate a permeable boundary between the assessment system and others, such as the teaching system, the faculty rewards system, or the systems of beliefs about pedagogy in the subject/area. People simultaneously belong to many of these systems and their actions in one are affected by their positions in others. Assessment cannot be understood well if it is viewed as an independent system. Much that is done by way of assessment derives from other systems that are resistant to attempts to change assessment practices. Questionable assessment practices may be sustained by activity systems that are not primarily about assessment: systems designed to instruct students in bodies of information may not work well as assessment systems. What is not shown in the figure is the suffusion of power through the assessment and other systems, which tends to maintain what is but which sometimes helps to create what is not.

The second theory, complexity theory, is presented as a good set of metaphors, rather than as a prescription (Thrift, 1999). Paradoxically, complexity theory illuminates the ways in which two complex systems that were almost-identical at start can develop very differently but it also says that once they have settled into a cycle it can be hard to make any significant, lasting change to the their functioning.

With established systems, the probability is that any change may wobble the dynamics but no more. Systems cycle around powerful attractors, with mighty forces being needed to change their dynamics. Figure 2 adds the idea that assessment systems become more complex as their objects — the students' achievements — move from information, through knowledge, to knowing. The more complex the system, the more uncertain the outcomes of any changes and the harder it is, then, to make a planned difference. Complex activity systems may not be reformable in quite the ways that rational change theories assume.

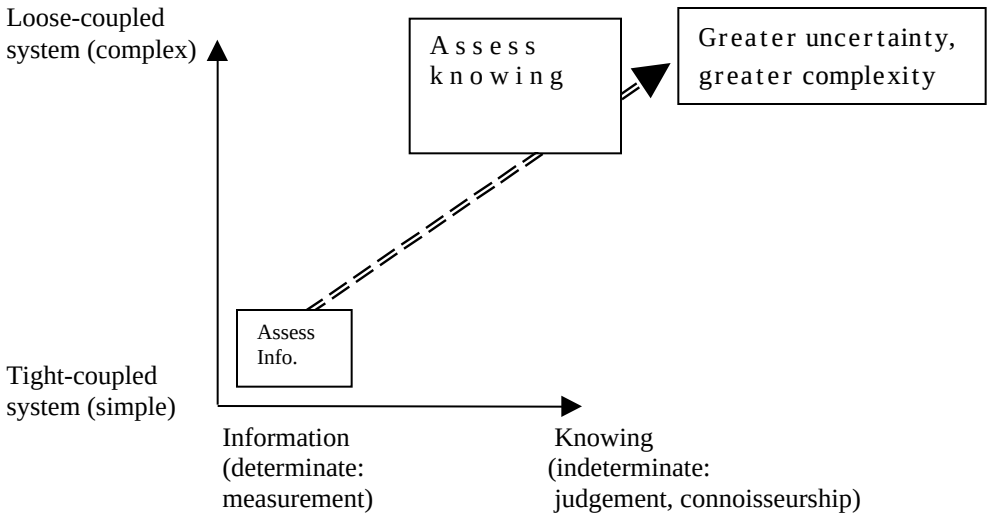

Figure 2. Assessment objects and system complexity.

The third theoretical set relates to communication. Although some researchers into information and communication theory are concerned with how to reproduce at one point either exactly or approximately a message selected at another point, others with a semantic interest do not see messages as objective packages of information to be posted and received intact (Fiske, 1990). Action has been taken to enhance the communicative value of higher education assessment data. In England, the Quality Assurance Agency (QAA) has sponsored the production of benchmarks, which identify some of the understandings and skills that characterise some subject areas. This is a step towards a common language and the expectation that assessment criteria should be explicit, used, shared with students and available to stakeholders complements it. New books on assessment continue to appear and it is customary for them to contain advice on assessing achievements that have often been missed by traditional methods, which goes some way to meeting the objection that assessment has had a narrowing effect. It is a moot point whether higher education's attempts to clarify summative assessments will resonate with the stakeholders who attribute meaning to them.

There is a deeper sense in which assessment practices may not be good for quality monitoring systems. Assessment is ultimately about judgement, not measurement.

Measurement is exact, assessment is not. Agreed, judgement and measurement can be synonyms when determinate achievements are concerned. In such cases, to draw on the language of figure 1, subjects and communities of practice are simply organised to use handy methods and the most basic of criteria to make low- or no-inference judgements of simple achievements. Lecture-and-test practices that are common in some subjects and some countries, where students face multiple choice tests of their short-term recall of information presented in lectures and texts, are examples of this convergence of judgement and measurement. Higher education, though, is essentially about less determinate achievements. As figure 2 suggests, when such achievements are involved, the amount of uncertainty in the system is much greater, it becomes more complex and measurement becomes a less realistic prospect. Some, such as Eisner (1985), have argued that connoisseurship is the best way to appraise these complex achievements, acknowledging that some achievements remain hard to appreciate and elude reliable discernment. The main objection is that this introduces excessive subjectivity. Attempts to reduce it by having well-trained experts using agreed criteria to make multiple observations in multiple contexts. Yet some achievements continue to elude recognition and, because these approaches are also expensive and demanding, there is a steady drift back downhill, in terms of figure 2, to the efficient measurement of determinate achievements.

There are all sorts of reasons why summative assessment data are not good performance indicators for quality monitoring. This section has also argued that it is hard to see how summative assessments could be sufficiently reformed to make a substantial difference. An alternative is to sidestep the problem of trying to shape recalcitrant reality into the form that PIs require. A four-fold scheme is proposed.

1. Information can be securely measured and reported. There is no need for change of practice here.
2. Some complex achievements, such as essay-writing, are comfortably assessed in local communities of practice. The actions that QAA has taken to encourage departments to share their local understandings and practices offer some prospect that there will be greater commensurability between judgements of some complex achievements made in different communities. This is a reduction of uncertainty about what judgements might mean, not its elimination.

These two measures involve little change to higher education assessment as it is. Degree awards and GPAs could rest on more-or-less the same basis as they do now. An implication of the argument developed in this paper is that it should be understood that they are incomplete and imperfect descriptions of local achievements and ought not to be used as PIs for comparative purposes. To rest on these two features would leave higher education with a tolerably reliable but very narrow assessment system. More is needed.

3. Students should make and support claims to those complex achievements that cannot be very fairly, reliably and affordably assessed. Just because higher education

institutions are not in a position to warrant achievements, it does not follow that students cannot claim them. It might follow that higher education institutions should take care to help students to learn the language in which claims could be couched, to see what might plausibly be claimed, how, and on what grounds. A job for quality monitoring would be auditing this education for claimsmaking.

4. Although the UK government's key skills agenda does not provide a secure basis for common, high stakes assessment, it does identify a growing consensus that well-educated graduates are likely to share some common achievements that are generally attractive to employers. If higher education institutions ought not to try and warrant the outcomes, they might consider certifying the processes. They might declare to stakeholders that students have engaged in a range of learning encounters that favour these sorts of learning achievements. Plainly a student making claims to achievements that can be related to appropriate documented learning engagements is more plausible than one who is unable to refer to sustained, progressive and varied encounters. Here, the task for quality monitoring systems would be to appreciate the learning engagements to which students in higher education institutions and on named programmes are entitled. It would be, as Eisner (1985) suggested, curriculum connoisseurship.

Given that assessment data are more liable to mislead than to lead, it is unwise to base IQM and EQM on them. An alternative is to appraise the quality of these systems for helping students to make good claims to complex achievements. Higher education institutions would then be accountable to employers, governments and stakeholders in general for the quality of processes that are commensurate with good, complex learning, not for the phantom achievements created by dubious summative assessment practices. Investment would be guided by internal quality enhancement processes, founded on principles of connoisseurship and continuous quality improvement. Insofar as they rely on assessment data, established EQM and IQM systems are outmoded. Type II approaches are called for.

References

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

Anderson, J. R., Greene, J. G., Reder, L. M. & Simon, H. A., 2000, 'Perspectives on learning thinking and activity', *Educational Researcher*, 29(4), pp. 11–13.

Boud, D., 1995, 'Assessment and learning: contradictory or complementary?' in Knight, P.T.,(Ed.). *Assessment for Learning in Higher Education*, 35–48 (London, Kogan Page).

Breland, H. M., 1999, 'From 2 to 3Rs: the expanding use of writing in admissions' in Messick, S. J. (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(2), pp. 141–53.
Cliff, N., 1996, *Ordinal Methods for Behavioral Data Analysis* (Mahwah NJ, Lawrence Erlbaum Associates).
Cuming, J. & Maxwell, G., 1999, 'Contextualising authentic assessment', *Assessment in Education* 6(2), pp. 177–94.
Dalziel, J., 1998, 'Using marks to assess student performance', *Assessment and Evaluation in Higher Education*, 23(4), 351–66.
Eisner, E., 1985, *The Educational Imagination* 2nd edition, (New York, Macmillan).
Engestrom, Y., 1990, *Learning, Working and Imagining: Twelve Studies in Activity Theory* (Helsinki, Orienta-Konsutit Oy).
Fiske, J., 1990, *Introduction to Communication Studies* 2nd edition (London, Routledge).
Fleming, N., 1999, 'Biases in marking students' written work', in Brown, S. & Glasner, A.. (Eds.) *Assessment Matters in Higher Education* pp. 83–92 (Buckingham, Society for Research into Higher Education & Open University Press).
Fullan, M., 1999, *Change Forces: The sequel*. (London, Falmer).
Greatorex, J., 1999, 'Generic descriptors: a health check', *Quality in Higher Education*, 5(2) pp. 155–66.
Hart-Landsberg, S., Braunger, J., Reder, S. & Cross, M., 1992, 'Learning the ropes: the social construction of work-based learning', *ERIC*, accession number 363726.
Harvey, L. & Green, D., 1995, *Employer Satisfaction* (Birmingham, CRQ).
Heywood, J., 2000, *Assessment in Higher Education*. (London, Jessica Kingsley Publishers.
Knight, P. T., 2000, 'The value of a programme-wide approach to assessment', *Assessment and Evaluation in Higher Education*, 25(3), pp. 237–51.
Knight, P. T. and Trowler, P. R., 2000, 'Academic work and quality', *Quality in Higher Education*, 6(2), pp. 109–14.
Linn, R., 2000, 'Assessments and accountability', *Educational Researcher*, 29(2),pp. 4–16.
Mitchell, J., 1997, 'Quantitative science and the definition of measurement in psychology', *British Journal of Psychology*, 88, pp. 355–83.
Price, M. & Rust, C., 1999, 'The experience of introducing a common criteria assessment grid across an academic department', *Quality in Higher Education*, 5(2) 133–44
Saunders, L., 1999, *'Value Added' Measurement of School Effectiveness: A critical review* (Windsor, NFER).
Thrift, N., 1999, 'The place of complexity', *Theory, Culture and Society*, 16(3), pp. 31–69.
Wolf, A. *et al*., 1997, *Assessment in Higher Education and the Role of 'Graduateness'* (London, HEQC).
Woolf, H. & Cooper, A. with others, 1999, 'Benchmarking academic standards in history: an empirical exercise', *Quality in Higher Education*, 5(2), pp. 145–54.
Yorke, M., 1999, 'Benchmarking academic standards in the UK', *Tertiary Education and Management*, 5 (1) 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(1), pp. 7–27.

P. T. Knight, Educational Research, Lancaster University,Lancaster, LA1 4YL,
p.knight@lancaster.ac.uk