The Role of Public Examinations in Defining and Monitoring Standards

MIKE CRESSWELL

Examination Standards and Educational Standards

I INTERPRET THE PHRASE *educational standards*, in its widest sense, to mean the quality of educational provision and I take it that interest in the monitoring of such standards is motivated by a desire for valid information upon which to base policies intended to improve that quality. Despite being concerned only with some of the objectives of education, public examinations play a major role in defining and monitoring educational standards because their results are often used (in, for example, school performance tables) as output measures for accountability purposes. Public examination standards therefore underpin much of the public debate about educational standards generally and, indeed, are themselves the focus of controversy. To be able to engage critically with the public debate on standards, it is therefore necessary to understand the nature of examination standards and the extent to which, given their nature, they can legitimately be used to draw conclusions about educational standards. In this paper, I attempt to illuminate some of these issues.

Defining Public Examination Standards

The word *standards* is a notoriously slippery one. One of its more misleading aspects, when it is used in the context of educational assessment, is the image which it conjures up of standard measures in the physical sciences. Somewhere in France, we are told, there was once a bar of metal which defined the exact length of a metre. Where is the educational equivalent kept?

Proceedings of the British Academy, **102**, 69–120.

Once it is asked, the very strangeness of this question gives pause for thought. But what is actually wrong with the idea of a standard measure of educational attainment? Why can't we, for example, simply keep somewhere a copy of a Grade A GCSE script which, upon inspection, will enable us to see exactly what *GCSE Grade A* represents? The problem, of course, lies not in keeping the script but in being able to see what it represents. In practical terms, despite the deep philosophical waters surrounding primary and secondary properties or the possibility of objective knowledge, physical properties such as length are directly observable. Thus, when we look at an object exemplifying a length of 1 metre, we can directly observe its length and could use it to measure directly the length of a second physical object. To do this, we would simply hold the example alongside the second object and make a visual comparison.

With the general example of an examination script, however, we are dealing with a human linguistic artefact and to identify the standard which it represents we must interpret the language which it contains. This is a major difference from the case of physical properties. Interpretation of written text is far removed from direct observation and the meaning which a particular text has to an individual reader depends not only upon the text itself but also upon what the reader brings to it (see, for example, Eagleton 1993). Thus our standard script will represent something different, to a greater or lesser extent, to each reader. Moreover, an examination script does not encapsulate the whole of what it exemplifies. Educational attainments are complex networks of knowledge, skills and understandings, not all of which will be assessed on any one occasion nor, therefore, exemplified in any particular script. Thus, it is necessary to infer from the archive script what represents the same standard in the unrepresented aspects of the attainment being assessed. This has particular importance if we wish to compare the archive script with a second script. Not only will the second script require interpretation, but in general it will cover a somewhat different subset of the attainment being assessed. When the scripts are compared, the comparison cannot therefore be direct but must be a comparison of the two inferred standards, each of which is based upon an interpretation by the observer.

A popular response to this difference between physical and educational measurement is to try to construct explicit descriptions of educational standards and I shall look at this approach in more detail shortly. For the moment, however, we simply need to note that such

descriptions are themselves linguistic artefacts and, in complex attainment domains, cannot be comprehensive. They therefore depend upon exactly similar processes of interpretation and inference to communicate standards as did our archive script. The upshot is that examination standards cannot be objective in the everyday sense in which standards relating to physical measurements are objective. The inevitable role of interpretation and inference in defining examination standards means that they are fundamentally subjective. That is to say, different observers will interpret the same student performance (or explicit description) differently, according to their different expectations and different notions of what constitutes attainment in the subject concerned.

To set an examination standard it is not therefore sufficient simply to identify a particular paradigmatic performance on a particular occasion, it is also necessary to address the question of how that performance should be interpreted and the implications which it has in terms of the wider range of attainments which make up the subject being examined. The detailed nature of *what* is measured by an examination is not obvious but a definition of it must be part of any useful definition of any particular examination standard.

In addition, the primary purpose of public examinations is to provide information for future meritocratic educational and vocational selection decisions (see Cresswell 1995 and 1997a, for detailed argument supporting this claim) and it is a feature of these selection processes that information from different examinations is combined and compared. As a result, it is essential, in terms of the notion of meritocratic fairness which underpins the use of examinations in selection, for all examinations of a particular family (e.g. GCSE) to report in terms of a common scale and for the same point on that scale to correspond to attainment with the same degree of merit for any examination in the family. That is, in examining jargon, the standards represented by the same grade from any examination in the family must be *comparable.* Other uses of examination results which treat grades from different examinations as interchangeable, such as school performance tables, also impose the same requirement.

Therefore, to define standards in a particular public examination in the way in which the term is normally understood, we must define:

1 what should be assessed;

2 the levels of attainment which are comparable to those represented by each grade in other examinations in the same family;

and to understand examination standards, we need to consider both of these aspects.

Turning, first, to the second aspect of examination standards, how can we determine whether the attainment required for the same grade is comparable in different examinations? To illustrate the difficulty of this question, take just a few specific examples: we require a Grade C in Mathematics to represent comparable attainment to a Grade C in Physics, a Grade C in English, a Grade C in French and a Grade C in Art. This requirement implies some way of making direct quantitative comparisons of candidates' attainments across disparate subjects. This is impossible because quantitative comparisons can only be made in terms of common features and the features which candidates' work in different subjects have in common are insufficiently relevant. That is to say, they refer to aspects of the work which are either irrelevant in terms of what should be assessed in one or both subjects (e.g. we could compare the physical properties, such as length, of candidates' answers but to what purpose?) or are only a part of what should be assessed in one or both subjects (e.g. quality of drawing, accuracy of spelling, arithmetic competence, and so on). To compare attainment in different subjects we are therefore left only with indirect bases for comparison, of which there are two: statistics and expert judgement, both of which I will look at in detail shortly.

First, however, we need briefly to consider the other aspect of examination standards: *what should be assessed?* Historically speaking, the curriculum in British schools evolved slowly, but continuously, to accommodate new approaches and developments. More recently, the curriculum has come under central control and short periods of comparative stability punctuated by imposed change may now be a more accurate description. In either case, however, the curriculum is not static and neither, therefore, is what is assessed in public examinations. The aspect of public examination standards which concerns what should be assessed therefore changes to reflect the current values of those in control of the curriculum or, where a national curriculum is not specified, those responsible for approving syllabuses. Moreover, it is not the case that such changes are necessarily small, as the following examples of changes in what is assessed illustrate: the introduction of 'modern' topics into mathematics syllabuses (and the accompanying removal of much Euclidean geometry) which began in the late 1960s, the removal during the 1970s, for ethical reasons, of dissection from

A-level Biology examinations; the change, for which the demands of international commerce were part of the public justification, to assessing communicative competence in modern foreign languages when GCSE was introduced in 1988; the continuous revision in what is assessed in examinations in computer studies to keep up with developing technology.

These examples illustrate the extent to which past examination standards have changed to reflect their evolving cultural context as well as overtly educational or pedagogical considerations. More subtle, but equally important, changes in what is assessed occur continuously in all school subjects, reflecting wider cultural changes in the way in which individual school subjects are perceived and contemporary expectations about students' achievements. Changes such as these, whether deliberate or evolutionary, mean that, as with the issue of comparability between subjects, the maintenance of comparable examination standards over time in what is nominally the same subject, actually requires quantitative comparisons to be made between qualitatively different attainments. Again, therefore, only indirect bases for comparison—statistics or expert judgement—can be used to maintain examination standards.[1] I will consider statistical approaches first.

Statistical Approaches

What is *comparable*, as applied to examination grade standards, normally taken to mean in statistical terms? In general, it is not taken to mean that an **individual** taking two comparable examinations should necessarily be awarded the same grade but is concerned, instead, with groups of candidates. For example, talking about the particular case of comparability between examining boards, the Forum on Comparability set up by the *Schools Council* (which had a general responsibility for British public examinations until the mid 1980s) said:

[1] In the evolutionary case, it is possible to argue that there is sufficient commonality between what will be assessed on any two adjacent occasions for practically useful, if slightly flawed, direct quantitative comparisons to be made between the attainments demonstrated by candidates. Even if this is exploited, however, there is an inherent *sorites* style paradox which means that, although standards are apparently maintained (within some close practical limit) between all adjacent years this does not guarantee their maintenance over long time periods. Most importantly, the accumulating qualitative changes in what is assessed and how it is evaluated force any check on the maintenance of standards over long periods to fall back on indirect means of comparison. I will return to this point later.

> . . . the expectation is that had a group of examinees followed another board's syllabus and taken its examination, they might reasonably be expected to have obtained the same average grade.' (Schools Council 1979)

The definition of comparability implicit in this quotation can usefully be elaborated to cover the comparability of every grade, by replacing the reference to the 'average grade' by a reference to the **distribution** of grades for the group of candidates and in this paper I shall take *statistical comparability* to mean the identity of grade distributions. With this stance, a variety of different operational definitions of comparability are generated, depending upon the approach adopted to the problem of defining groups of candidates for which it is reasonable to expect identical grade distributions (see Cresswell 1996, for more details). However, my arguments below about the limitations of statistical definitions of comparable standards apply in general and do not depend crucially upon any particular definition.

Elsewhere (Cresswell 1996, 1997a, Goldstein and Cresswell 1996), I have set out in detail the conceptual problems surrounding statistical approaches to defining and maintaining comparable public examination standards in different examinations. Here, I will briefly summarise the two major ones. First, candidate attainment is affected by many variables, such as quality of teaching, student motivation and so on, as well as by examinations and their associated syllabuses. To illustrate the problem which follows from this, suppose that the standards in two examinations are set so as to produce identical grade distributions giving statistically defined comparability but that, subsequently, extensive guidance on effective teaching approaches for one of the syllabuses is provided. If this has the (presumably, desired) effect of improving the quality of teaching of the syllabus, the grade distributions of the two examinations will now differ and, according to our statistical definition, they will no longer set comparable standards *even if the question papers remain unchanged and are marked and graded identically.* This is not consistent with the way in which the notion of examination standards is normally interpreted and those who wish to use statistical definitions of comparable standards must either accept that they are open to continual revision[2] or argue that the standards originally set are to be

[2] This is, of course, catastrophic for those whose motive for using statistical approaches is a desire to make examination standards objective so that they can monitor changes in candidates' attainment over time. The continual revision of reference groups retains a sort of objectivity, but the price paid is the loss of any possibility of monitoring over time—the very reason for wanting objectivity in the first place. I shall return to this point later.

preferred, despite being based upon a comparison made on an arbitrary historical occasion.

The second major theoretical problem with statistical definitions of comparable standards concerns identifiable subgroups of candidates. Even when two grade distributions are identical for the whole groups of students whose attainment they describe, they need not be identical for well-defined subgroups within those groups. For example, the differences between boys' and girls' performances in GCSE examinations are well known and depend, at least in part, on the assessment techniques used (see, for example, Murphy 1982, Cresswell 1990) so that, if two examinations using different techniques have identical grade distributions for boys and girls combined, they will not necessarily have identical grade distributions for the boys and girls considered separately. Other reasons may produce similar disparities and GCSE English and Mathematics examinations illustrate the sorts of effects which occur in practice. Over the entire 16+ age cohort, the boys' grades in these two subjects are similarly distributed, but the girls' grade distributions differ substantially (Newton 1997a). It is unclear how this sort of effect can be accommodated if comparable standards are to be defined in terms of the identity of grade distributions. Are GCSE English and Mathematics standards comparable for boys but not for girls, *even though both boys' and girls' work is marked and graded identically*?[3] In the light of this, those wishing to define examination standards statistically have two choices again: re-define examination standards in a way which does not reflect their normal meaning (though it is not at all apparent, in this case, how this could be done) or defend the choice of an arbitrarily constituted group of candidates with which they define their standards.

In practice, even if the two theoretical problems discussed above are side-stepped by making arbitrary choices about the groups of candidates to be used and the occasions when comparable standards are to be defined in terms of the identity of grade distributions, major difficulties remain. Grade distributions reflect the attainments of the candidates who take the examinations and these attainments are the result of the interaction of many different variables. Systematic differences between the self-selected groups of candidates who take different examinations are therefore to be expected and so, in practice, it becomes necessary to attempt to control for such differences in the attainment of the candidates taking the different

[3] Ignoring any unintended gender bias which, in a recent study (Baird, 1998), was found not to be significant.

examinations before convincing statistically comparable standards can be established. Either an independent measure of attainment needs to be employed or indirect control of attainment can be attempted through school and student variables which influence it. Elsewhere (Cresswell 1996), I have reviewed in detail the various approaches of this type which have been tried in the past; here it is sufficient simply to note that the choice of relevant control variables inevitably involves value judgements and that different choices will lead to different operational definitions of comparable standards (for an example and an excellent discussion of the issues see Baird and Jones 1998).

The single most important lesson to draw from this discussion is that the use of statistical definitions of comparable standards does not, as is sometimes thought, lead to some sort of objectivity in examination standards. This is an illusion created by the technical layer which statistical approaches insert between the standards themselves and the, often implicit, value judgements which must underpin operational choices of historical baselines, specific reference groups of candidates and, in any practical system, appropriate control variables.

Judgemental Approaches

The other approach which can be used to set examination standards is to use expert judgement to define the levels of performance in a particular examination which represent standards comparable to those represented by the grades awarded via other examinations in the same family. This approach is theoretically coherent, provided that the judgements involved are accepted as *value* judgements, because, as most philosophers of value argue, value judgements do not ascribe properties to the objects being judged.[4] As a result, setting standards via judgements of the relative value of candidates' attainments in different subjects does not involve making quantitative comparisons between qualitatively differing attainments. Thus, using expert judgement to define examination standards is theoretically justified, provided that the judges are seen as expressing their subjective view about the value of the candidates' attainments and not as identifying some objective property of the attainments which, by virtue of their expertise, they can recognise.

[4] For general discussion on this point from a range of philosophical perspectives see, for example, Ayer (1946), Fogelin (1967), and Billington (1988). French *et al.* (1987) argued the same case specifically for judgements of educational attainment.

Moreover, the judgemental approach to defining examination standards is not vulnerable to the same theoretical problems as the use of statistical definitions since it does not make the standards depend directly upon the performance of candidates. Thus, for example, changing differences between the proportions of boys and girls who meet a judgementally determined standard have no implications for the coherence of the standard itself, nor does an internal inconsistency arise if the outcomes in one examination change relative to another. In theory, such changes can be interpreted directly as changes in the attainment of the candidates, as valued by the judges.

Of course, to argue that setting examination standards is not a matter of identifying performances which meet some set of objective criteria but is a process of subjective judgement more akin to evaluating a work of art or literature, raises immediate questions about its acceptability. Value judgements are usually seen as subjective and irrational as a consequence of the affective component which they contain (as in: *I don't know much about Art, but I know what I like*). However, this says more about our everyday notions of expertise and rationality than it says about value judgements. It has long been recognised that, although value judgements cannot be facts amenable to empirical verification or epistemological justification, they can, nonetheless, be the results of a rational process and be supported by reasons (Fogelin 1967, Beardsley 1981) if not pure deductive or inductive reasoning (Best 1985). Moreover, there is now considerable neurological evidence to suggest that feeling is an essential, if unrecognised, part of what we would normally class as rational thought (Damasio 1995) so that deductive or inductive reasoning is rarely, if ever, the sole basis for the decisions which human beings make. Value judgements are no different. Although they are subjective, they can be based upon reasoned argument, are not necessarily simply emotional or intuitive responses and should not, therefore, be assumed to be irrational, necessarily capricious or, indeed, unreliable (in the sense of being difficult to replicate). In fact, ten years ago, Frances Good and I obtained levels of agreement equivalent to marking reliabilities above 0.95 between different groups of examiners judging the quality of GCSE scripts in History, French and Physics (Good and Cresswell 1988a).

However, the reliability and rationality of value judgements is not always easily accepted and, in Britain in the last 15 years or so, there have been many attempts to create more objective judgemental systems for setting examination standards via systems of explicit criteria and aggregation rules. If, as I have just argued, setting examination

standards is an evaluative process, it seems reasonable at first sight, as Christie and Forrest (1981) proposed, to try to identify the criteria which judges use, and how they use them, when they decide the relative merits of candidates' work and thus the grades which they should be awarded. I call this approach *strong criterion-referencing* and give an extensive discussion of the ways in which it differs from more conventional criterion-referencing in Cresswell (1997a). Implicit within it is an over-simplified model of the process of human judgement which has considerable intuitive appeal.

This implicit model sees the process of setting standards as one which involves judging the relative merits of students' work (for example, their examination scripts) by using a fixed set of *criteria* which epitomise work of varying standards in each of a number of different *dimensions* of attainment. In conventional practice, the criteria are *tacit*, in that they exist only in judges' heads, but strong criterion-referencing assumes them to be capable of explicit expression, given sufficient introspection and linguistic ingenuity. Having decided the relative quality of the work in terms of each dimension of attainment, the judge is then thought to synthesise an overall judgement of each script as a whole by combining the judgements on the separate dimensions using a process of rational thought. It is assumed in strong criterion-referencing that there are implicit high level computational procedures which produce this synthesis and, again, that these procedures are capable of being made explicit, given sufficient introspection. This view of judgement is thus an essentially mechanical and dualist one, in which the judges first formulate a description of each candidate's work in terms of criteria on several different dimensions and then consider their description, presumably in some sort of *Cartesian theatre* (Dennett 1993) in their heads, so that they can apply high level computational rules to determine its overall value.

I call this view of judgement the *Cartesian Computer Model* and it might help to make it clearer if I outline how it was operationalised in one recent attempt to make it concrete: the approach initially used in England and Wales for National Curriculum Assessment at the end of Key Stage 3. Here, the different *Attainment Targets* of National Curriculum subjects represented the different dimensions, each consisting of 10 levels of increasing competence (since reduced to 8) defined by 10 sets of explicit criteria called *Statements of Attainment* (now re-written in whole sentences and called *Level Descriptions*). Candidates were assigned to a level on each Attainment Target separately and then

explicit aggregation rules were used to produce a single overall level for each candidate from his or her separate levels on the Attainment Targets.

Readers who doubt the intuitive appeal of the Cartesian Computer Model, and suspect me of erecting a straw man, should reflect upon its remarkable tenacity, in association with the strong criterion-referencing meme, in the face of repeated failure. In the 1980s, there was a considerable amount of work done in Britain on the notion of *grade criteria* for public examinations (Hadfield, 1980 [quoted in Christie and Forrest, 1981]; Orr and Nuttall, 1983; SEC, 1984; Forrest and Orr, 1984; Orr and Forrest, 1984; Bardell *et al* 1984; SEC 1985; Long, 1985; SEC, 1986; SEC,1987). This work assumed the Cartesian Computer model, being based on the view that standards can be encapsulated in written criteria which prescribe the level of attainment required to justify the award of a particular grade. However, although it was possible to write standard-setting grade criteria (either *ab initio* as in the original SEC work or as a result of perusing candidates' scripts, as in later work), in use they proved not to apply to some candidates' performances which were awarded the grades in question by conventional procedures. Gerry Forrest and his colleagues (Forrest and Orr 1984; Orr and Forrest 1984, Bardell *et al.* 1984), in their general conclusions, published in the reports of all their studies, observed:

> performances in existing examinations that would result in the award of particular grades may not qualify for those grades if the criteria considered relevant were applied.

Since then, further attempts to make the Cartesian Computer model of judgement operational, most notably the early National Curriculum assessments, have also failed. The most recent failure is visible in the apparently endless process of revision of the assessment models used for General National Vocational Qualifications and Key Skills. In fact, no strongly criterion-referenced public assessment system has ever been made to work successfully, given that success means replacing holistic value judgements of quality with an equivalent set of explicit criteria and aggregation rules. It is, of course, possible to challenge this definition of success but the consequence of using explicit criteria and aggregation rules which are not successful in these terms is that the examinations concerned do not meet commonly accepted requirements related to fairness (see Cresswell 1987, Cresswell and Houston 1991, Cresswell 1994 and Cresswell 1997b for more on this aspect). In any

case, the intention of work on public examination grade criteria has always been to make the standards represented by the existing holistically judged grades explicit, not to change them.

A fundamental problem with grading systems based upon explicit criteria and aggregation rules is that they view the task of judges as involving the identification of a set of observable well-defined qualities in candidates' work. In fact, as Sadler (1987) has argued, the criteria which are appropriate for any particular educational value judgement differ according to the nature of the work being evaluated. This has long been recognised as a feature of judgements of value in other fields (see, for example, Aldrich 1963). Indeed, the presence of the same quality may be reason to value one object highly but not another, because other features of the second object change the value attached to the quality in question (Pole 1961). Modern students of literary criticism also argue that the nature of the work modifies the criteria relevant to its evaluation (Eagleton 1993). Moreover, work on the psychology of evaluative judgement is consistent with these views. Eiser (1990) reviews the relevant empirical psychological evidence, and shows that it strongly suggests that a person presented with new information searches for relevant conceptual categories with which to encode it, starting with those immediately accessible in memory and continuing until sufficient (according to some internal criterion) relevant ones are found. In addition, the 30 year-long failure of rule-based artificial intelligence to reproduce human expertise outside very tightly systematised domains is now beginning to be seen as evidence that human expertise does not involve the rapid application of very complex high-level rules. Authors such as Dreyfus (1992) and Devlin (1997) argue that what distinguishes an expert is precisely that he or she has transcended such rules and developed holistic skills specific to his or her area of expertise.[5]

[5] Although it may be possible to reproduce expert behaviour in some domains using systems of catalogued knowledge and high-level rules, the formation of evaluative judgements of educational attainment looks unlikely to be such a domain. Such judgements require the implicit formation of statements like 'This script has enough of the properties required for a Grade A.' According to Copeland (1993) the set of valid sentences of this type is formally *undecidable*—that is to say, there is no algorithmic process which can always determine if a sentence is a member of this set—it follows that no computational rule, however complex it is, can be guaranteed to produce the correct overall evaluation from all possible quantitative judgements on various dimensions unless the set of those quantitative judgements is finite, which seems unlikely. It is also worth noting that recent practical attempts to use Artificial Intelligence-style techniques to categorise students' answers to free-response test questions use a brute force statistical modelling approach which the researchers concerned (Burstein *et al.* 1997) suggest is best used *alongside* but not *in place of* human judgement.

Thus, the human evaluative process does not involve the application of a limited and well-defined set of criteria which have been determined beforehand to be relevant to all objects of a particular class, and nor is it exhaustive or rule-based. The above mutually supporting philosophical arguments and empirical evidence reveal the fundamental reason why strong criterion-referencing, involving the application of explicit written criteria according to the Cartesian Computer Model of judgement, does not, and cannot, replicate value judgements made by suitably qualified judges.[6]

A more suitable model for the process used by examiners evaluating students' examination work is the *Multiple Zen Drafts Model.* This model of judgement sees the judges as engaged in a constant process of evaluation and re-evaluation as they read the candidate's work. There is no need for a pre-existing set of evaluative criteria and therefore no set of computational rules for reaching an overall judgement. However, there is normally reasonable agreement between different judges about which features of candidates' work are likely to be relevant to its evaluation. The evaluation is direct and immediate in the somewhat metaphysical way celebrated by Pirsig (1974) in *Zen and the Art of Motorcycle Maintenance* but is continuously open to revision (re-drafting—hence *multiple drafts*) as the judge reads more of the work. The judge reads and re-reads the work until his or her evaluation stabilises. (See Dennett 1993 for an extensive discussion of *multiple drafts* as a model of consciousness in general).

The Multiple Zen Drafts Model is consistent with modern understandings of the nature of critical evaluation, particularly with *Reader Response (or Reception) Theory.* I can do no better to explain Multiple Zen Drafts further than quote part of Eagleton's (1993) description of the process of reading:

> . . . the reader will bring to the work certain 'pre-understandings', a dim context of beliefs and expectations within which the work's various features will be assessed. As the reading process proceeds, however, these expectations

[6] This is not, however, to say that concise statements cannot be constructed which, in Wilmut and Rose's (1989) terms, 'convey the flavour' of a grade. Such *grade descriptions* have been included in GCSE syllabuses, for example, since 1988 and describe a paradigmatic attainment worthy of the grade, rather than the attainment of every candidate awarded the grade. Although, for the reasons given, grade descriptions cannot be used as criteria for objectively judging the attainment of all candidates, they can form a useful focus for building consensus among examiners about the sorts of qualities attached to scripts which they judge worthy of a particular grade and can thereby help to facilitate agreement between different examiners' qualitative judgements.

> will themselves be modified by what we learn, and the hermeneutical circle—moving from part to whole and back to part—will begin to revolve. . . . What we have learnt on page one will fade and become 'foreshortened' in memory, perhaps to be radically qualified by what we learn later. Reading is not a straightforward linear movement, a merely cumulative affair: our initial speculations generate a frame of reference within which to interpret what comes next, but what comes next may retrospectively transform our original understanding . . .
>
> (Eagleton 1993: 77)

There is a certain irony in the fact that Reader Response Theory, as such, was developing during the 1970s and early 1980s, at the very time when educational assessment was most obsessed with the essentially mechanical Cartesian Computer Model of evaluation enshrined in the notion of criterion-referencing and the associated attempts to render evaluative criteria completely explicit. Few public examination awarders would fail to recognise the similarity between the process described by Eagleton and their own experience of judging candidates' scripts.

The Nature of Examination Standards

The preceding analysis leads to only one conclusion: examination standards are not, *and cannot be made*, objective[7]—they are social constructs. The need for all definitions of examination standards to define *what* must be assessed requires judgements of a non-technical nature to be made, relating to issues such as the wider social value of different educational attainments. Less obviously, even within a single school subject, what is to be assessed and the perceived value of particular levels of attainment change over time to reflect contemporary concerns so that standards only have meaning within their own social context, localised in time and space.[8] As a result the bases upon which candidates separated in time can be assessed are different, making direct quantitative comparisons of their performance invalid (see also Goldstein 1983 and Cresswell 1997a). Only indirect statistical comparisons or qualitative value judgements can provide a theoretically valid basis for the comparison of performances from different occasions in domains which change over

[7] This is not to say that the assessments of individual examination candidates cannot be 'objective' in the sense of being free of significant biases and reliable; indeed, recent research on these matters is reasonably reassuring (see Baird 1998 and Newton 1996).

[8] Interested readers can consult Cresswell (1996 and 1997a) or Wiliam (1996a and 1996b) for more detailed discussion of the socially constructed nature of examination standards and its implications.

time and both of these approaches involve subjectivity, whether in the process of judging candidates' work or in the arbitrary choice of statistical reference groups. It follows that the objective measurement of quantitative changes in educational standards is impossible.

However, these are theoretical concerns and a tough-minded pragmatist might well argue that they are somewhat precious. Such a critic needs to know why, if subjectivity is simply acknowledged as inevitable, public examination results cannot still be used as one fallible quantitative measure of changes in educational standards in the same way as the retail prices index, the composition of which also varies over time, is used as one measure of domestic inflation. After all, he or she might argue, the examination boards claim to maintain the same standard from one year to the next, albeit by the use of qualitative judgement and/or statistics based upon reasonable but arbitrary decisions. Surely, therefore, comparisons of the statistics of examination outcomes over long periods of time must tell us something about changing overall standards of attainment. To respond to this challenge, in the next section I take a practical look at the processes by which comparable public examination standards are maintained from year to year and draw out the implications of those processes for the interpretation of changes in examination statistics over time.

Maintaining Examination Standards in Practice

In the British public examination system a new examination is set each year on any particular syllabus and there is therefore a need to establish a new pass mark for each grade (these marks are called the *grade boundaries* in examining parlance) on each successive version of the examination. Grade boundaries cannot simply be carried forward unchanged from one examination to the next because the papers inevitably vary somewhat in difficulty from year to year, requiring compensating changes in the marks required for the award of each grade if the standards of attainment demanded for those grades are to be comparable between years. To determine the new positions of the grade boundaries, awarding meetings are held in which senior examiners make judgements about the quality of work of sample candidates' scripts and combine the qualitative data produced by this process with statistical evidence to arrive at final recommendations for the positions of the grade boundaries. The process is specified in outline in QCA (1998) and described in detail in Cresswell (1997a). Its use of

qualitative judgement and indirect statistical evidence is consistent with the theoretical analysis I gave in the preceding sections. Although, in practice, both qualitative judgement and statistical data are used to maintain examination standards, it is instructive to consider these two sources of evidence separately.

Using Expert Qualitative Judgement

Turning to qualitative judgement first, how effective is it for maintaining examination standards? That is to say, can examiners, by scrutinising candidates' answers to examination papers, identify the mark which corresponds to the same standard of attainment as a given grade boundary mark in the preceding version of the examination? If they can, then any changes in the examination outcomes which follow from their qualitative judgements will arise from changes in the attainment of the candidates between years, not from any differences in difficulty which there may be between the two years' examinations. On the other hand, differences in outcome which could be shown to be related to the examination itself would imply that the judgemental process had failed to take into account those differences in difficulty. Thus, considering changes in the examination outcomes which would follow from only the judgemental part of the awarding process, casts light on the adequacy of qualitative judgement *per se* as a method of maintaining standards. The question is: can the scale and nature of such changes in outcomes reasonably be explained only by changes in the attainment of the candidates or is there reason to believe that effects due to the examination and its awarding are also present? This section addresses this question.

However, it is essential to be clear about the nature of the investigations which I report here. In the absence of independent information about the attainments of the candidates, it is impossible to disentangle the effects of candidate attainment and examination difficulty within any observed changes in a *particular* examination's statistics. This section therefore considers the balance of probabilities, based upon an investigation of many different examinations. This essentially statistical approach enables the existence to be demonstrated, with a high level of probability, of changes in examination outcomes which are due to the examinations, rather than the candidates. It does *not*, however, enable the particular examinations in which such effects operated to be identified.

Table 1 shows the changes produced between two successive examination occasions in the cumulative percentages of candidates

Table 1. Comparisons between the outcomes in two successive examinations derived from purely judgemental data for a sample of A–level examinations with more than 500 candidates.

	Number of candidates in		Change in cumulative percentage at		
Subject	Year 1	Year 2	Grade A	Grade B	Grade E
Accounting	6775	6637	−1.4	−3.9	−0.1
Applied Mathematics	1359	1237	−6.5	−3.4	0.1
Biology I	4152	5159	−4.2	−9.4	−12.4
Biology II	2195	2464	−5.0	−8.4	−12.7
Business Studies	11206	12477	1.3	3.2	6.2
Chemistry	3902	4139	−2.7	−5.8	−8.0
Communication Studies	3945	4565	2.5	6.4	2.1
Computing	3445	3294	−2.6	−4.2	−0.7
Constitutional Law	701	669	0.0	0.8	13.6
Economic & Social History	928	1066	−1.2	−2.6	−4.4
Economics	12056	11913	0.1	−0.8	−0.9
English I (Language & Literature)	11196	12186	−0.3	−2.1	−0.9
English II (Literature)	4401	3680	−4.8	−11.4	0.1
English III (Literature Alternative)	10061	13929	4.2	0.9	−3.6
Environmental Science	562	794	−0.2	0.4	1.4
French	4492	5324	1.2	5.4	5.1
General Studies	1141	1256	2.8	6.6	8.9
Geography	2813	3026	1.4	8.8	9.0
German	1706	2140	−1.6	−7.0	−10.0
Government & Politics	1658	1835	0.1	−2.9	1.8
History	5234	5894	0.0	−2.2	−4.2
History (Alternative)	2125	2397	−0.6	−1.6	−3.4
History of Art	1233	1226	1.9	8.5	6.0
Human Biology	2832	3300	0.0	0.1	0.4
Law	3747	4166	−4.2	−3.8	−0.1
Philosophy	476	699	−3.8	−4.1	−15.3
Photography	1335	1307	0.9	0.4	0.0
Physics	7412	7423	0.0	−2.8	−5.2
Physical Education	222	630	0.4	1.6	−1.1
Psychology	8504	10120	1.1	1.2	4.0
Pure & Applied Mathematics	6718	6647	−0.3	0.6	−0.2
Pure Mathematics	3219	3137	−13.2	−15.7	−4.4
Pure Mathematics & Statistics	5464	5432	0.4	1.6	−1.4
Sociology	19789	17222	0.5	−2.7	0.7
Spanish	773	928	−4.0	−11.8	−10.7
Sport Studies	477	749	−0.3	−1.8	2.4
Statistics	1939	1778	2.0	9.8	9.9
Theatre Studies	4749	5634	−0.9	−0.5	−1.4

awarded each of three grades, when the grade boundaries were fixed purely by a process of qualitative judgement of candidates' work. The data relate to grades A, B and E in a broad range of A-Level examinations which attracted over 500 candidates in the year in which the data were collected. Some examinations exhibit small changes in overall outcomes between the two years, some examinations exhibit large ones. One point immediately worth making is that there is no obvious relationship between the nature of the subjects and the scale of the changes.

In public examination awarding procedures, examiners are asked to explain any large changes in the proportions of candidates who would be awarded each grade as a result of their qualitative judgements. Three possible explanations, one or more of which will be offered by the awarders in any particular subject, are usually proposed. In *Explanation Number 1*, the examiners simply argue that the entire group of candidates, as such, are better (or worse) than the previous year's group. In *Explanation Number 2* they go a little deeper and refer to any changes there may have been in the relative proportions of candidates entered by different types of centres or in the gender balance within the entry. (Descriptive information about the composition of the entry is routinely available to awarding meetings in terms of these two variables.) *Explanation Number 3* concerns the case where the number of candidates entering for the examination has changed considerably. It may then be suggested by the examiners that the candidates who have been gained or lost, as a group, are better (or worse) than the rest of the candidates.

In the following sections, I examine each of the examiners' three explanations, in the context of the data as shown in Table 1.

Explanation Number 1: As a whole, the candidates as a group are simply better (or worse)

By itself, this is not an explanation at all for a change in the proportions of candidates in each grade, it is simply a restatement of the implications of the grade boundary judgements in a different form. Since it is offered by the same individuals who have made the judgements which it purports to explain, it is not independent corroboration of the results of those judgements and cannot, logically, explain them.

However, this explanation is interesting because of the implied models which are held by those who proffer it. The explanation could be based upon the notion that systematic overall attainment changes are to be expected because of changing educational policy and practice

or other external factors; or it could be the reflection of an implicit assumption that some variation is to be expected between the results of adjacent years' candidates simply because they are different groups of students; or it could be referring to both of these potential causes for variations in examination outcomes. I will consider the second cause first. Given that it is clearly a possibility, the obvious question to ask is: how large are the random variations which can be expected in the statistics of public examination results? This question effectively views each year's candidates for a particular examination as a sample from the population of all candidates who take that examination over its lifetime. Posed like this, it is essentially a question for sampling theory.

If a few plausible assumptions are made about the nature of examination candidate populations and the randomness of any particular year's entry (see Cresswell 1997a), then it is straightforward to evaluate the size of the differences in examination results statistics which might be expected as a result of chance differences between successive years' entries. The standard test for the significance of differences between proportions in large samples can be used to evaluate the statistic, z, which is theoretically normally distributed with a mean of 0 and variance of 1. If this test is applied to the data producing Table 1 (see Cresswell 1997a, for details), the results summarised in Figure 1 are obtained. Clearly, the differences in outcomes between the two years cannot reasonably be viewed as the results of random variations between successive groups of candidates.

It follows that some effect other than straightforward sampling error is operating in many of the cases in Table 1. It might be argued that the prior ability or motivation of the groups of candidates entered for the different examinations changed significantly between Years 1 and 2 for some systematic reasons. Possible causes of such a change might include, for example, widespread medical factors or social ones such as growing fear of unemployment, although such extrinsic factors seem unlikely to affect different school subjects differentially, as they would have to do to explain the data in Figure 1. Alternatively, in the terms of Explanation Number 1, the causes of the changes in outcomes between the two years must be improvement or deterioration in the quality of educational provision in the subjects concerned. Thus, to evaluate Explanation Number 1, the key question is whether changes in the overall ability or motivation of candidates or in educational provision could be the causes of the significant changes in examination outcome illustrated as shown in Figure 1.

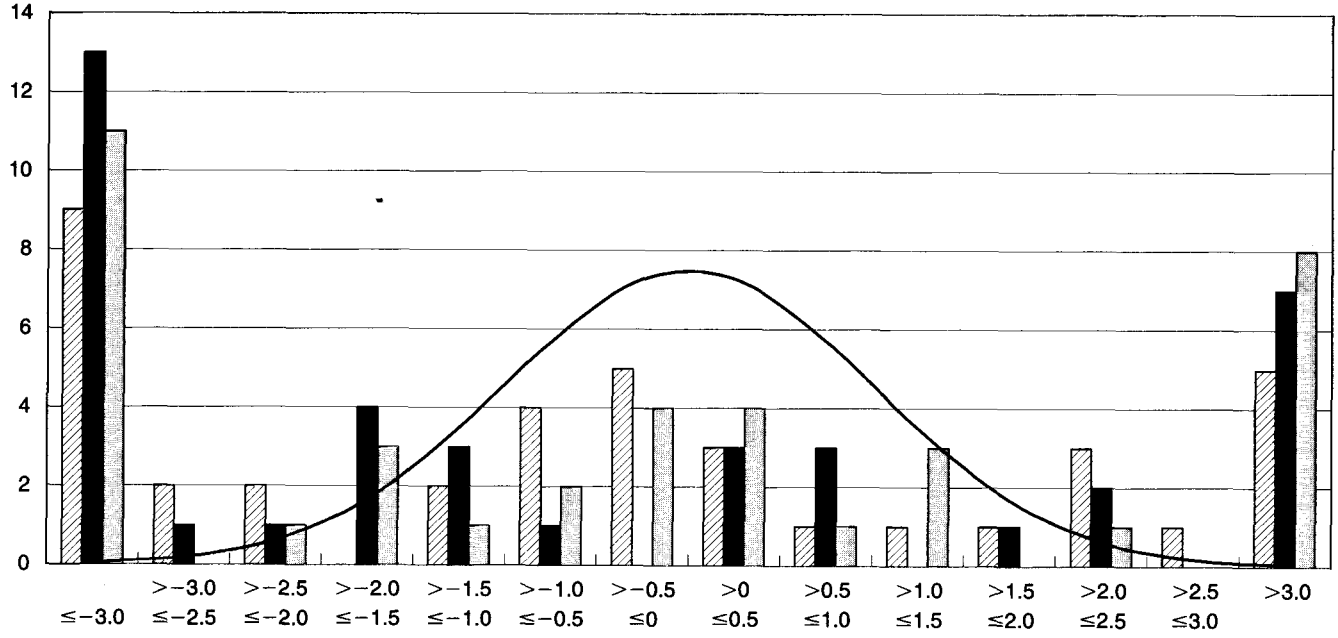

Figure 1. Distribution of z statistics for differences between judgemental outcomes in two years for a sample of A-level examinations. (▨) A, (■) B, (□) E, (——) normal.

With the data available, this question cannot be unequivocally answered. However, evidence relevant to it can be obtained by considering the changes in outcomes which would follow from the judgements for subgroups of candidates whose origins are different and whose educational provision is differently organised. This has been done for the examinations in Table 1 by looking separately at the changes in outcomes between Years 1 and 2 for UK candidates from schools, UK candidates from further education colleges and overseas candidates (the detailed data are contained in Cresswell 1997a). In all, there are 67 cases in Table 1 where there is a significant difference between Years 1 and 2 in the cumulative proportion of all candidates at a key grade boundary. In 47 of these cases, a change in the same direction occurs for all three subgroups of candidates and, in the remaining 20 cases, such a change occurs for 2 out of the three subgroups.

Moreover, if the same analysis is done for the changes in outcomes in the same subjects between Year 1 and the preceding year (the full data are, again, given in Cresswell 1997a) and the results are compared with those illustrated in Figure 1, it is clear that the annual changes in outcome between these adjacent pairs of years are little related to each other—see Figure 2. It is difficult to identify any plausible extrinsic factors or educational mechanisms which could not only differentially affect overall attainment in different subjects on a global scale so markedly, but also produce effects which vary so much from one year's

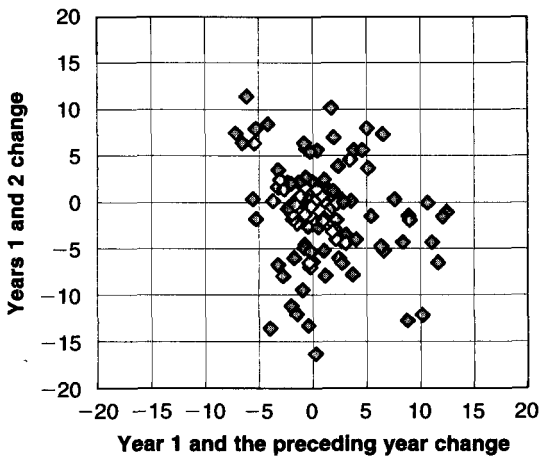

Figure 2. Relationship between z statistics of changes in outcomes for Year 1 and the preceding year versus Years 1 and 2.

cohort of candidates to another. On balance, therefore, Explanation Number 1 appears insufficient to explain the scale of changes observed, between years 1 and 2, in the proportions of candidates awarded each grade if qualitative judgements of candidates' work *are the only data* used to maintain standards. Any mechanism capable of generating the changes observed would have to operate through something which all subgroups of candidates within an annual cohort have in common but which changes annually. The examination itself is the only known factor which meets these requirements.

Explanation Number 2: The balance of centre types and/or genders has changed

This explanation seems at first sight to be a plausible one. Clearly, if there are differences in attainment between different subgroups of candidates, then variations in the relative proportions of these subgroups will lead to changes in the overall proportions of candidates awarded each grade. Is this effect sufficient to explain the differences observed in Table 1?

To explore this question, the grade distributions for subgroups of candidates in Year 1 (see Cresswell 1997a) were combined, re-weighted in such a way as to reflect the relative proportions of each subgroup in Year 2, as follows:

$$P'_x = \sum_j P_{jx} \cdot s'_j$$

where P'_x is the predicted proportion of Year 2 candidates exceeding the boundary for Grade x,

P_{jx} is the proportion of Year 1 candidates in Subgroup j exceeding the boundary for Grade x

and s'_j is the proportion of candidates in Subgroup j in Year 2

The changes in outcome predicted in this way were then compared with the actual changes in overall grade distribution between the two years. Clearly, any differences between the actual changes in outcomes and the ones predicted by re-weighting indicate discrepancies which cannot be accounted for by changes in the composition of the entry between the two years, at least with respect to the subgroups referred to in Explanation Number 2. The observed differences are not only very much larger than those predicted but also uncorrelated with them (see Cresswell 1997a, for the detailed analysis and a demonstration that, in general, realistic changes in subgroup distributions are unlikely to produce large

changes in overall examination outcomes). Explanation Number 2 is not, therefore sufficient to account, in general, for the scale of changes observed in the outcomes of successive years' examinations if qualitative judgement *alone* is used in an attempt to maintain grade standards.

Explanation Number 3: This year's new (missing) candidates are better (or worse) than the rest

This explanation, which is sometimes offered by examiners when the number of candidates entering for an examination has grown or shrunk considerably, is essentially a special case of Explanation Number 2 in which the new (or missing) candidates are thought of as a subgroup of candidates with zero incidence in the previous (or current) year. As a result, the plausibility of Explanation Number 3 as a sufficient explanation for observed changes in grade outcomes is similar to that of Explanation Number 2. Only if the entry for an examination grows or shrinks substantially as a result of many different centres making similar changes to their entry policies and entering candidates from a different range of attainment, can Explanation Number 3 account for large changes in the proportions of candidates awarded each grade. However, as Table 1 exemplifies, year-on-year changes in the number of candidates entering for an examination are rarely large in proportion to the existing entry.

The relationship between examiners' judgements and the statistics of the candidates' marks

It appears very probable, therefore, that the differences in outcomes reported in Table 1 are due, at least in part, to fluctuations in the standards represented by the examiners' qualitative judgements. As was noted earlier, it is not possible, in the absence of independent assessments of the candidates' achievements, to *prove* this conclusion in any *particular* case, nor to estimate the precise size of the discrepancies which occur. However, it is possible to establish upper bound estimates for the movements in grade boundaries[9] which would have been required to set standards in Year 2 which were comparable to those in Year 1. If it is assumed that the attainments of the Year 2 candidates were distributed identically to those of the Year 1 candidates, then changes in the distributions of marks between the two years

[9] By convention, the *grade boundary* is the lowest mark on the examination awarded to work which merits the grade in question.

can be interpreted as indicating changes in the difficulty of the question papers and/or changes in the severity of the marking process. Since the purpose of awarding comparable standards is to make adjustments to grade boundaries which compensate for such changes, estimates of the positions of the Year 2 grade boundaries can then be obtained by scaling the grade boundaries used in Year 1 in accordance with the means and standard deviations of the Years 1 and 2 mark scales, as follows:

$$B'_x = \frac{(B_x - m_y)}{s_y} \cdot s_{y+1} + m_{y+1}$$

where B'_x is the new position of the boundary B_x for Grade x
m_y is the mean score in year y
and s_y is the standard deviation of scores in year y

The results of doing this have been compared with the positions of the grade boundaries based solely upon examiner judgement in Year 2, producing Figure 3 which plots the actual movements of the grade boundaries against the movements predicted on the basis of the changes in the mark statistics. (Two of the examinations in Table 1 have been excluded because their maximum mark changed substantially between Years 1 and 2, producing misleadingly extreme movements.) It can be seen that there is a fairly strong relationship between the predicted and actual grade boundary movements but that, on average, the size of the actual movements is about 0.4 of the size of the predicted ones. From this analysis, it appears that the examiners correctly identified the direction of the changes required but, given the present assumption that the candidates were of comparable achievement in the two years, failed to take sufficient account of the change in difficulty of the examination papers and/or their marking.

Although the assumption upon which the present analysis is based is exactly that, an assumption, it is important to note that, in 81 (77 per cent) of cases, the qualitative judgements moved the boundaries in the *direction* implied by the mark statistics, if not to the *extent*. Thus, in these cases, the judgements confirm that, to some extent at least, the mark statistics reflect changes in the difficulty of the examinations. However, there is no reason to believe that any change there might be in the attainment of the candidates from one year to the next is not independent of any change in the difficulty of the examination which occurs. Therefore, if the reasonable assumption is made that the candi-

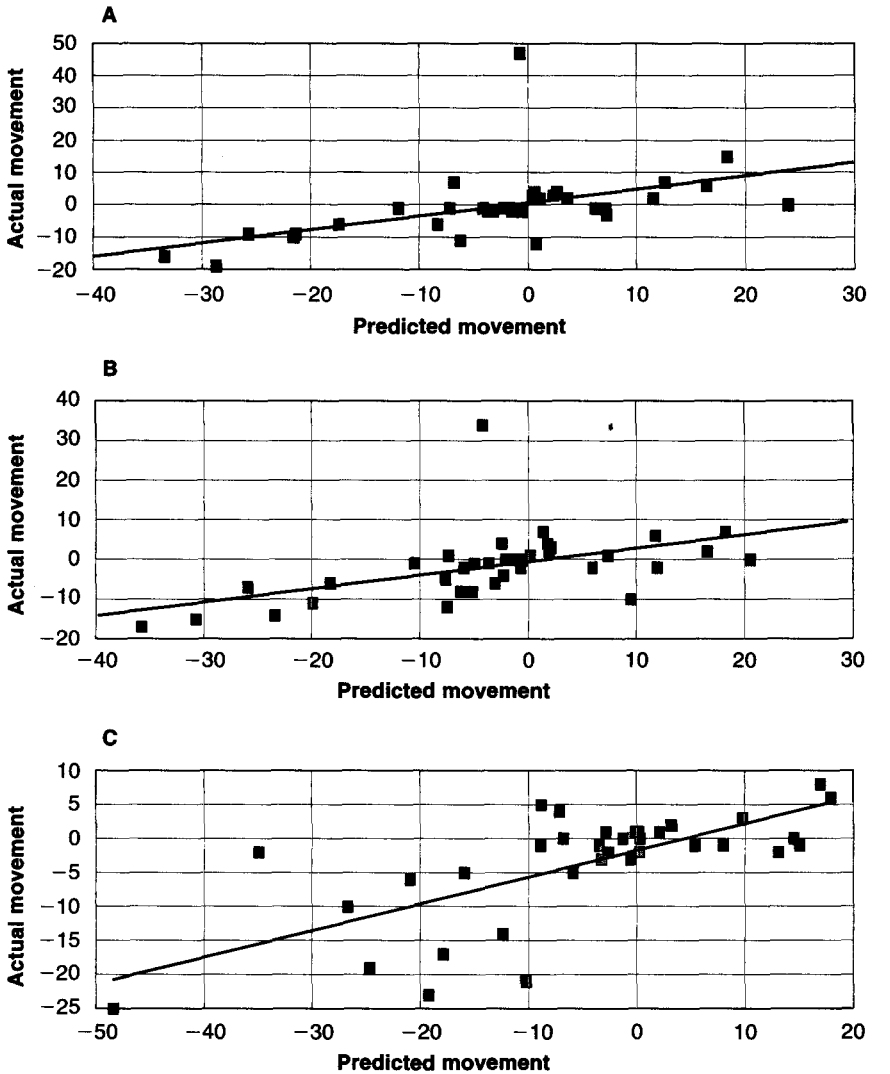

Figure 3. Actual movement of Grade boundaries in Year 2, against movement predicted from change in mark statistics.

A.	Grade A boundary	$y = 0.4156x + 0.6181$	$R^2 = 0.2504$
B.	Grade B boundary	$y = 0.3407x - 0.5835$	$R^2 = 0.2545$
E.	Grade E boundary	$y = 0.3928x - 1.7727$	$R^2 = 0.4898$

dates are equally likely to be slightly better or slightly worse from one year to the next, the actual movement of the grade boundaries should be less than that predicted from the change in mark statistics in 50 per cent of cases and greater than the predicted change in the remaining 50

per cent. However, of the 81 cases where the mark statistics and judges agree on the direction of the move, the actual move is less than the predicted move in 57. Using the binomial distribution, the two-tailed probability of this (or a more extreme value) occurring by chance is easily shown to be less than 0.001. This strongly suggests that, for some of these examinations at least, the examiners' qualitative judgements took insufficient account of changes in the difficulty of the examination papers and/or their marking.

This conclusion is entirely consistent with Frances Good and my (Good and Cresswell 1988a and 1988b) earlier experimental result that judges tend towards relative severity when setting grade boundaries on harder papers within tiered examinations. It seems reasonable to conclude that qualitative judgement alone is inadequate as a method of maintaining examination grade standards from year to year because it does not take sufficient account of changes in the difficulty of successive year's examinations. Some of the reasons for this relate to psychological and sociological aspects of the judgemental process which have recently been studied in detail (see Murphy *et al.* 1996, and Cresswell 1997a). Earlier, I set out at some length the reasons why attempts to reduce the process of qualitative judgement to a more mechanical activity involving explicit criteria and rules for reaching overall judgements based upon them—what I call strong criterion-referencing—does not offer a solution to this problem and has failed every time it has been attempted. The only alternative method of maintaining examination standards is therefore the use of statistical evidence either instead of, or alongside, qualitative judgement.

Using Statistics

There are several different statistical approaches, of varying technical sophistication, which could be used, by themselves, as the basis for maintaining examination standards from year to year (see Cresswell 1996 for a review). In practice, however, the procedures which the examining boards' use combine statistical data with qualitative judgement, as required by the regulator's *Code of Practice* (QCA 1998). In essence, the statistical parts of the procedures assume that large changes in outcomes from one year to the next are implausible for examinations with reasonably large entries from a stable group of schools. Analyses of the results in the previous and current year of only those centres which enter candidates in both years are sometimes used when there are

reasons to doubt this assumption but the use of these analyses does not change the argument in this section. Similarly, analyses which consider the effects of changes in the types of centres entering candidates or which use considerations of value-added or any other of the approaches which I reviewed in Cresswell (1996) may be used but, again, without changing the essential argument. In the interests of clarity, I shall therefore refer throughout this section only to the conventional approach in which similar statistical outcomes are expected from year to year for any particular examination.

In normal British examining practice, the combination of statistical and judgemental data is, itself, a judgemental process (QCA 1998). Formal methods such as, for example, the use of Bayes' theorem are not used (see Cresswell 1997a for further discussion). However, there is no doubt that the statistical data have a significant influence on the examination standards which are set and, thus, on the overall examination outcomes. Figures 4 and 5 show the equivalent data to Figures 1 and 3 when current awarding procedures, combining judgement and statistical data are used (see Cresswell 1997a, for more detail).

Figure 4 brings us to the crux of the issue over the maintenance of examination standards and the use of examination results to monitor educational standards. Is the slight upward movement in the outcomes the result of improved attainment in some examinations or of a slight tendency to leniency in the statistically informed standard-setting process? Or, more realistically, perhaps we should ask: to what *extent* is the slight upward movement in the outcomes due to improved attainment and to what *extent* is it due to the awarding process itself? Unfortunately but inevitably, there are no data available from the examinations themselves which enable this question to be answered.

What is clear, is that the active use, during the annual process of setting examination standards, of assumptions of statistical continuity between examination results in successive years means that the statistics of examination outcomes do not necessarily reflect changes in the attainment of candidates which occur over time. The very process of awarding removes, to some unknown extent, the effects of any such changes from the examination statistics. This makes it, at the least, potentially misleading to try to use public examination results for the long term monitoring of educational standards. Unfortunately, as the data reported earlier showed, if assumptions of statistical continuity are not made, the alternative unsupported qualitative judgements of students' work are insufficiently stable for their results to be interpreted

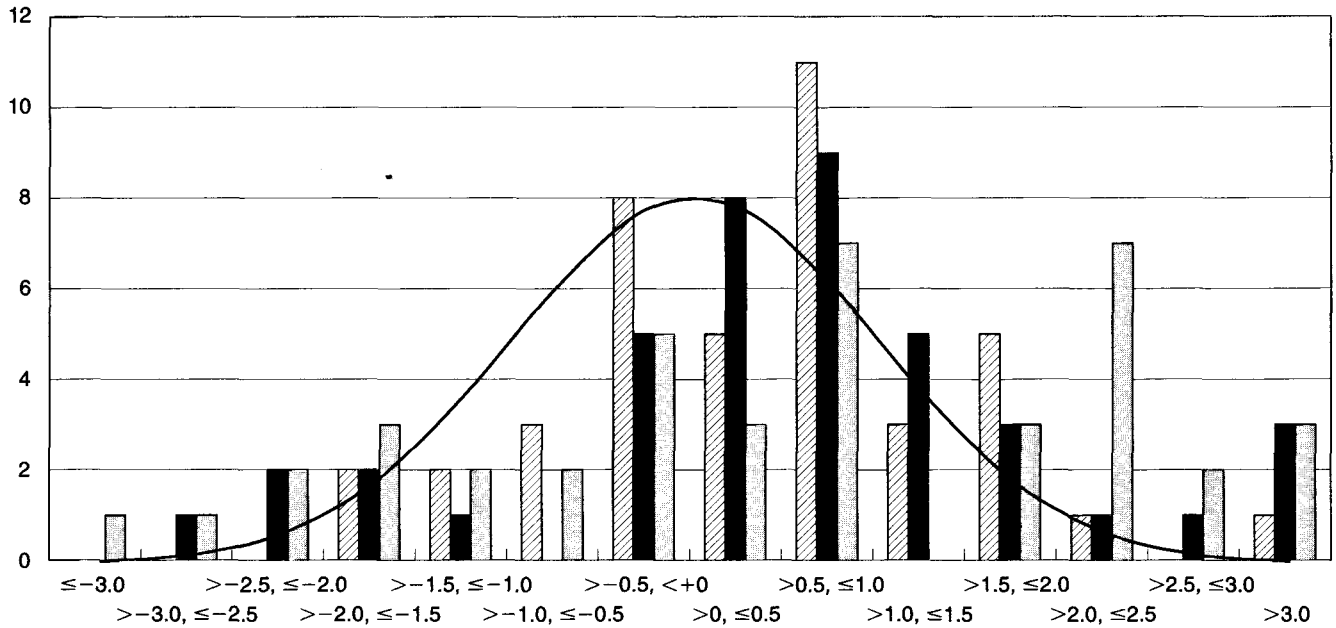

Figure 4. Distribution of z statistics for differences between outcomes for each key grade in two years—statistical data used. (▨) A, (■) B, (□) E, (——) normal.

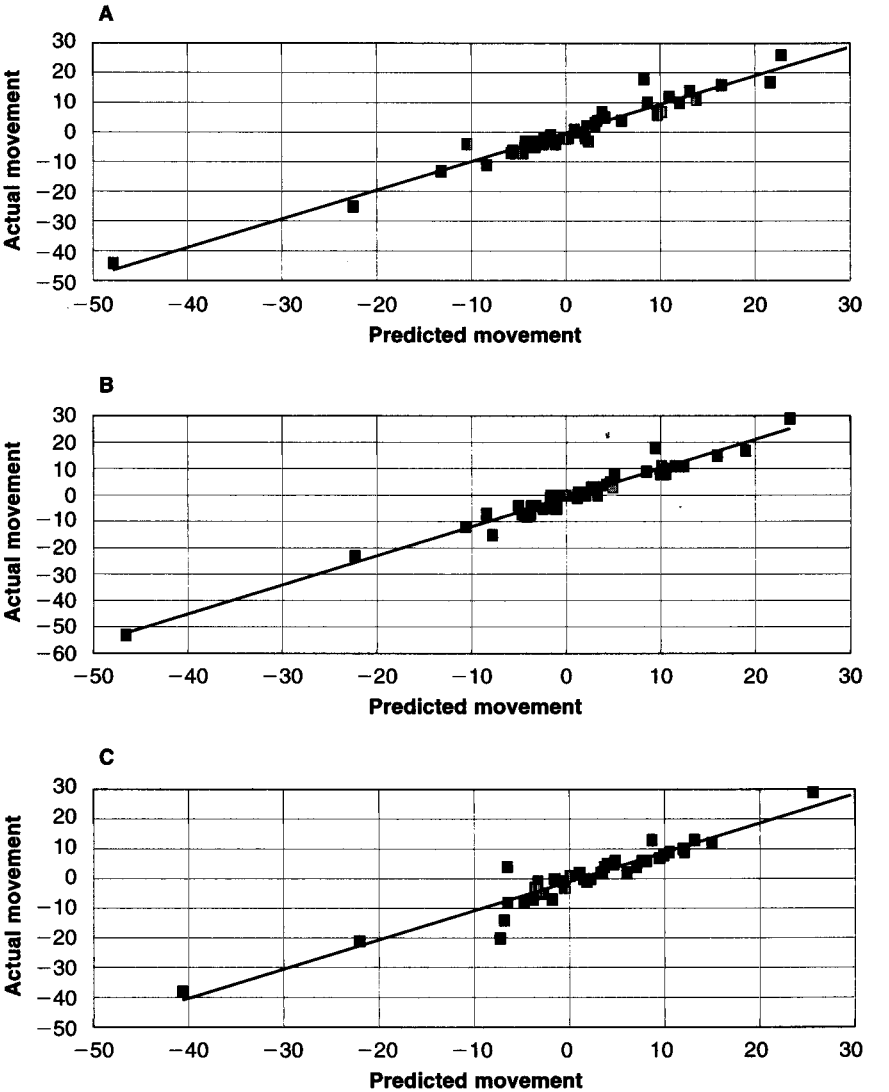

Figure 5. Actual movement of Grade boundaries in Year 2, against movement predicted from change in mark statistics—statistical data used.

A.	Grade A boundary	$y = 0.9643x - 0.3875$	$R^2 = 0.9423$
B.	Grade B boundary	$y = 1.1048x - 0.9194$	$R^2 = 0.9629$
E.	Grade E boundary	$y = 0.9822x - 0.9577$	$R^2 = 0.8966$

as reliable indicators of changing standards of attainment among examination candidates.

Moreover, the problem would not be solved (though it might be illuminated) by measures such as pre-testing public examination questions. The very high-stakes nature of public examinations for the candidates themselves means that problems of motivation and preparedness afflict non-operational pre-tests and, in any case, the need for security of the examinations means that papers could not be pre-tested in their entirety. Here, in conditions which reflect the historic purpose of public examinations—to award qualifications—lie some of the practical reasons why it is problematic to interpret the data which they produce in terms of changes in educational standards over time.

Similarly, the use of statistically informed qualitative judgement to set annual standards is entirely appropriate for the primary purpose of public examinations—the provision of qualifications. Sensible assumptions of statistical continuity between adjacent years, taking into account any known changes in the provenance of the candidates, are likely to hold reasonably well because major changes in overall educational standards are unlikely to be rapid. Moreover, although the use of statistical data in the annual awarding process is prey to all the problems discussed earlier in relation to arbitrary choice of reference groups, differential performance by well defined sub-groups and so on, the size of the effects concerned is unlikely to be large between adjacent years. In particular, a new reference group (the previous year's candidates) is used on each occasion. The implication is that standards can be maintained to a reasonably close tolerance between any two adjacent years but that this does not guarantee that standards set, say, 15 years apart are comparable, though they may be. Because the selection processes in which examination results are most important usually involve individuals who were awarded their grades within a few years of each other, the primary role of examinations as qualifications is reasonably secure. The use of examination results as accurate indicators of quantitative changes in educational standards over long time periods is, however, perilous.

Since the assumption of statistical continuity which is made during examination awarding tends to reduce the scale of changes in outcomes which would otherwise result, it could be argued that changes in public examination outcomes reflect long term changes in overall attainment but tend simply to understate them. This argument, of course, depends crucially upon an assumption that there is no long term bias in the process of qualitative judgement of candidates' work and here the

socially constructed nature of examination standards is important. The expectations of the examiners who judge the quality of candidates' work are rooted in their professional experience of their own students' attainments, discussion with their colleagues, contact with current educational thinking and, indeed, current political issues and social mores. Clearly, such expectations will not be static but will evolve to reflect changing social and educational influences. Once again, from the long-standing use of public examination grades in educational and vocational selection, it seems reasonable to infer that such dynamic norms, subject to statistical controls, provide standards which are stable enough to underpin useful individual qualifications. It does not follow, however, that they offer a sufficiently constant basis for measuring changes in attainment over long periods of time.

Public Examination Results as Monitors of Educational Standards

To emphasise the points from the preceding section in concrete terms, in this section I will briefly consider the interpretation of some recent historic data.

Figure 6 shows how success rates changed between 1985 and 1995 in a particular assessment. However, before I say whether Figure 6 relates to one or more GCSE or GCE examinations, the reader is invited to

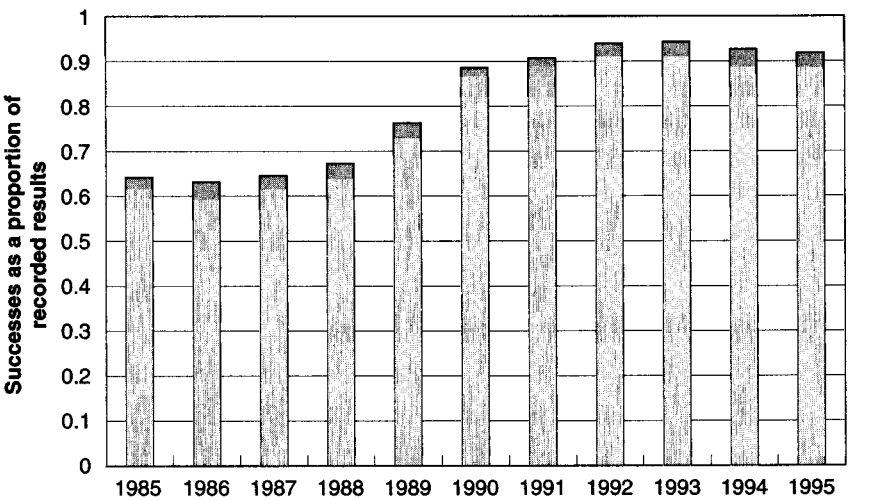

Figure 6. Success rate variations between 1985 and 1995.

pause and consider the question: does the graph indicate a rise in the standards of attainment of the candidates between 1988 and 1992 or a reduction in the assessment demands? Particularly adventurous readers might like to consider whether both effects took place and, if so, what their relative contributions to the overall pattern were.

In fact, Figure 6 shows the proportion of people reaching the summit of Mount Everest, expressed as a proportion of those who reached the summit or died on the mountain.[10] Since Mount Everest has not shrunk significantly in recent years, the interpretation of Figure 6 presumably has to be that people have got better at climbing it. Does this mean that climbing Mount Everest has become less demanding? In one sense, the answer is probably *yes*. For example, better equipment and more thorough preparation based upon the experience of earlier expeditions is likely to have contributed to the improved success rate. Here, then, is a further question to reflect upon: do such improvements represent an improvement in mountaineering standards or not?

Figure 7 shows the national percentage of girls in the Year 11 cohort who were awarded a Grade C or better in GCSE/O-level/CSE English between 1985 and 1995. Hopefully, it is now clear why interpretation of this graph in terms of either falling examination standards or rising attainment is problematic. In the Mount Everest example, appeal to the common human experience that mountains do not normally change height on short time scales enabled us to rule out one interpretation of the data but in the case of examination outcomes there is no such common experience. Thus, the two sides in the annual argument which greets the publication of public examination results about whether educational standards are rising or examination standards are falling are defined by their preconceptions about the relative likelihood of improving educational standards on the one hand or changing examination standards on the other. Since the examination data cannot, themselves, provide evidence one way or the other, they contribute nothing to the debate except a focus for argument. Either interpretation can be defended but neither can be proven without recourse to other information which is both sparse and, itself, controversial. It follows that serious attempts to monitor educational standards quantitatively

[10] Data from the New York Times (as reported at http: //everest.mountainzone.com/98/facts.html), un-amended except for the use of a rolling average to remove annual fluctuations caused by the relatively small number of attempts made on the summit of Mount Everest in any one year.

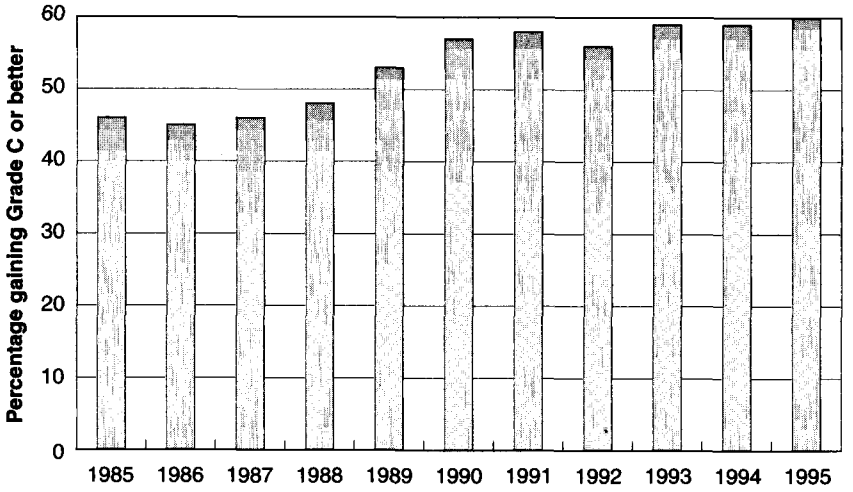

Figure 7. Year 11 girls' pass rates in GCSE/O-level/CSE English 1985–1995. Data from *GCSE Results Analysis: an analysis of the 1985 GCSE results and trensds over time*, published in London by the School Curriculum and Assessment Authority in 1996. (Note that the apparent drop in 1992 coincides with the use of a different original data source for 1992 onwards.)

must use information other than the statistics of public examination results.

Further questions arise in the light of Figure 8 which shows the national percentage of boys in the Year 11 cohort who were awarded a Grade C or better in GCSE/O-level/CSE English between 1985 and 1995. Comparison of Figures 7 and 8 shows that the improvement in boys' reported results was substantially less than that for girls over the time period shown. This raises several new interesting questions such as:

- Are GCSE English examinations increasingly biased in favour of girls or are educational standards for girls improving at a faster rate than those for boys or are there social phenomena leading to a growing gap between the performance of boys and girls or is there some other explanation?
- If the overall pass rate for boys and girls combined had been kept constant from 1985 to 1995, the boys results would have declined, does this mean that boys were really getting worse at English while the girls got better?

The examination results themselves can shed no light on the answers to these questions but it seems worth noting that those who want to

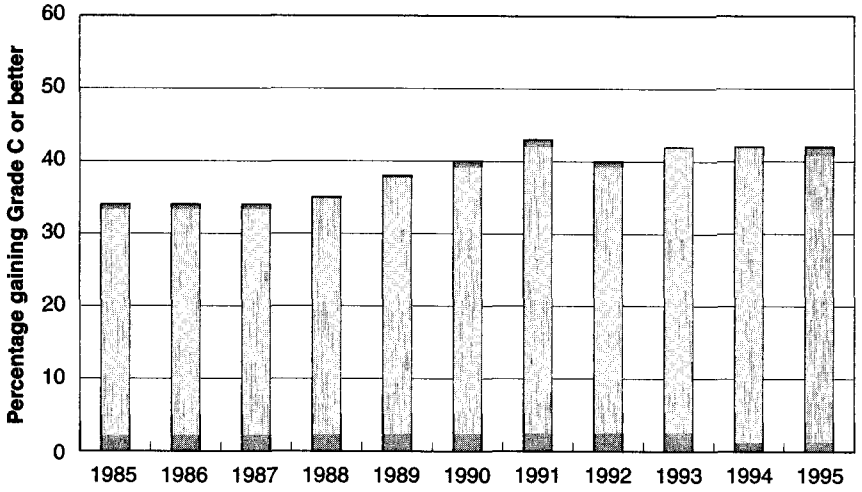

Figure 8. Year 11 boys' pass rates in GCSE/O-level/CSE English 1085–1995. Data again taken from *GCSE Results Analysis: an analysis of the 1985 GCSE results and trensds over time.* (Note that, as in Figure 7, the apparent drop in 1992 coincides with the use of a different original data source for 1992 onwards.)

interpret the data purely in terms of falling examination standards must be able to explain how those standards have fallen more for girls than for boys, even though they have taken identical examinations. In any case, there are many explanations for changes in examination candidates' results, relating to demographic, social and administrative variables, which mean that interpretation of examination statistics *per se* in terms of the quality of education delivered by the school system would still be impossible (see Newton 1997b, for an excellent review of these issues).

What contribution can public examinations make to monitoring standards?

Finally, I want to move away from what some might consider a rather negative analysis to a more positive perspective. I have argued that the objective measurement of changes in educational standards over time is impossible in theory. I have also argued that, even if theoretical concerns are set aside in a pragmatic search for useful, if fallible, quantitative measures of general educational standards, public examinations are not the answer. In particular the procedures used to set standards in public examinations have been developed to produce useful individual qualifications but their results cannot be reliably interpreted as quantitative indicators of long term changes in educational standards.

However, that is not to say that public examinations cannot provide valuable evidence of a qualitative kind about changing educational standards. In particular, past examination syllabuses, question papers, marking schemes and candidates' scripts provide a resource which could be used far more than it is to study the ways in which educational objectives and expectations and, indeed, the attainments of candidates have changed over time. The result of such studies would, of course, be relatively complex qualitative descriptions of change, rather than single headline statistics and such studies would say little about the current effectiveness of educational institutions compared with the effectiveness of comparable institutions in the past. However, rich qualitative descriptions of changing expectations and attainments would more properly reflect the true nature of educational standards and the complexity of the questions implicit in any attempt to monitor them over periods of more than a year or two.

Descriptive monitoring of this kind would, of course, reflect the values of those carrying out the studies and would not be objective but I have already argued that all approaches to monitoring standards involve subjectivity. The analyses would inevitably reflect current values so, if pressed to comparative conclusions, could only ever say that we, today, with our current values, prefer the past or prefer the present. However, are not our present values and beliefs a sufficient basis for action? Indeed, what other basis would we ever want to use? Perhaps the most dangerous aspect of the enterprise is that the result might be used to provide apparently 'scientific' justification for the pursuit of self-serving agendas, based upon idiosyncratic values, which could be claimed by those involved to be as legitimate a basis for action as any other set of values.

The best defence against this possibility is not, however, to deny that qualitative, and necessarily subjective, comparisons over time have any value nor to pretend that objective quantitative comparisons are possible. It is to confront the basic questions: What is the purpose of education? Is it economic success for our country; economic success for our children as individuals; the transmission of our culture; the fullest possible development of each individual's character and talents; or all of these,[11] in which case how do they interact and what relative emphasis should they be given? From these matters we might then move on to look in detail at the nature and degree of the attainments we wish

[11] No doubt readers can think of several other candidates.

our children to acquire and begin to develop useful answers to questions about the educational standards to which we aspire. Scrutiny of examination scripts would then be one source of subjective qualitative evidence as to whether our aspirations were being met. However, for the theoretical and practical reasons which I have set out in this paper, the statistics of public examination outcomes cannot, and do not, provide objective or unequivocal quantitative measures of temporal changes in educational standards.

Discussion

John Gray

Introduction

The view that performance in public examinations provides some kind of objective yardstick for judging educational standards is widely held. Examination boards are seen as the main keepers of these standards. Mike Cresswell's observation, therefore, that there is in reality no equivalent of the French metre rule against which to judge educational standards is not simply refreshing in its tacit acknowledgement that the annual debates about standards are flimsily-based. Crucially, he raises questions about how rigorous the practices so-called 'keepers of the standards' adopt to ensure their maintenance actually are and, equally importantly, about what further activities might reasonably be developed with the same aim in mind. His distinctive contribution is to make plain the assumptions underlying current practices—the picture he reveals is simultaneously familiar and disturbing (Cresswell 1998).

Much of Cresswell's paper focuses specifically on 'the maintenance of comparable exam standards between years' and the 'related matter of interpreting changes in the statistics of public examination results over time'. Both questions are central to concerns about maintaining standards. To come to conclusions about such issues exam boards use a mixture of both 'qualitative judgement and statistical data'. Somewhat unusually Cresswell offers accounts of both. His treatment of the statistical issues informing the maintenance of standards is, however, a good deal fuller than his account of the use of qualitative data. This imbalance certainly reflects the weight of the research literature and is,

perhaps, inevitable but makes one wonder how exactly potentially conflicting pieces of evidence from the two sources are reconciled.

Maintaining Standards as Social Practice

In the absence of the equivalent of physical yardsticks Cresswell argues that maintaining educational standards is best viewed as a social practice. This is a powerful insight whose implications he explores in some depth. Standards are not maintained by reference to some external yardstick(s) but through the social construction of inter-subjective agreements amongst those most closely involved—and principally the examiners themselves.

'Shared agreements' play a central role in contemporary philosophical thinking about objectivity and the nature of truth. However, a philosophical perspective is clearly not sufficient and needs to be accompanied by a rigorous analysis of the social practices employed to maintain standards. Who is recruited to be an examiner and who is promoted within the system to more senior positions? How are new recruits inducted and existing examiners refreshed? Crucially, since the system depends on the belief not merely that shared judgements are possible but that they actually exist, what does the evidence suggest about the extent of initial agreement amongst examiners? And, given these starting points, to what extent do the various processes to which exam boards then subject their examiners increase the likelihood of so-called agreements? In philosophical terms such judgements are likely to be most 'objective' when they are reached by 'unforced' processes.

Few of these issues have, to date, been researched in any depth. Only occasional glimpses are available about how examining committees conduct their business. There is, however, a more fundamental hurdle to be overcome. The idea that an exam board 'maintains standards' over time is almost certainly so integral to its culture and functioning that it seems doubtful whether any individual board could reach anything other than an essentially positive assessment of its own contribution. Some intensive analysis of how a board responded during those episodes when there was some possibility of 'breaking frame' would be necessary to cast light on this issue. But, frustratingly, such instances would themselves probably be dismissed as 'exceptional' by those most closely involved.

In sum, whilst there might be some value in exploring those occasions where an exam board was not maintaining standards from year to year (or

in danger of not doing so) the low incidence of such events would scarcely constitute independent evidence. As Cresswell argues, if one accepts that examination standards are 'socially constructed' then their legitimacy is based on the willingness of 'students, parents, teachers, employers and policy-makers' to 'accept the competence of the judges'. He suggests that 'the competence of those who set standards is more likely to be accepted if the procedures used are transparent and public knowledge' but the maintenance of trust is the key. It must be recognised that one of the reasons why there is still relatively little research on the maintenance of standards as a social practice is that such evidence might itself damage public confidence rather than reinforce it.

Statistical Contributions to Maintaining Standards

Over the last decade, as Cresswell demonstrates, exam boards have adopted increasingly sophisticated statistical procedures to explore standards-related issues. Their basic analytical armoury, however, has remained largely unchanged—namely what happened last year. There are good reasons for this of course. Last year's results are 'secure' in several respects: the pool of candidates taking any particular exam, for example, is likely to be fairly similar; the majority of examiners will probably have been involved the previous year and probably have set similar kinds of questions; the curriculum being examined can be assumed to bear comparison over such a short passage of time; and, perhaps as importantly bearing in mind the need to inspire public confidence, in the great majority of cases the published results will have been accepted subject only to a few marginal appeals. The cumulative effect of this reasoning is to reinforce the view that what happened last year provides a good starting point.

In the circumstances it is hardly surprising that what seems like a valid set of assumptions become translated into a set of working practices. Cresswell suggests that there are just three arguments that any particular group of examiners can employ to explain and justify changes from last year's results. These are that:

1 as a whole, the candidates as a group are simply better (or worse) than last year's;

2 the balance of centre types and/or genders has changed; or

3 this year's new (or missing) candidates are better (or worse) than the rest.

What is striking about all three explanations is how little additional evidence examiners are offered to inform their judgements. No evidence is offered, for example, with respect to the first explanation about whether this year's candidates have performed better/worse than last year's on some other (related) measure of performance on some previous occasion. The assumption is that they are basically the same unless there is powerful evidence to suggest otherwise. However, from Cresswell's account it would appear that no routine attempts are made to check out whether this assumption is valid.

Whilst examiners are offered evidence on the changing balance of centre types and gendered entries, the assumption is again that variations in performance between years will be small. the this year/last year frame of reference reinforces this impression. Cresswell comments that 'about 75% of centres entering candidates in any one year also entered candidates for the same examination in the previous year' and sees this as supporting judgements about the stability of the pool of candidates. Yet it is obvious that across three years the changes could be more substantial. If, as he suggests, around 75% of the centres one year are present the next then it is possible that in the third year only 56% of the original centres will remain.

Similar considerations could apply to the third explanation which Cresswell sees as a variant of the second. Examiners seem to make essentially optimistic assumptions about the extent of stability amongst pools of candidates drawn from different years. In the past, when only a minority of pupils were entered for any public examinations, the assumption that the pool of comparably-qualified candidates might have expanded could have been tenable; this seems less likely, however, when almost all pupils are entered for some examinations. In the circumstances it is hardly surprising if much of the examiners' practice seems like informed guesswork reinforced by (some) statistical insights.

Further Factors Contributing to Change

Over the last decade the proportions of young people securing 5 or more grade A–C passes in the 16+ examinations has been rising year-on-year at historically unprecedented rates. How, some critics ask, can these changes be squared with the maintenance of standards? The introduction of the new GCSE examination in 1988 (replacing the old bi-partite system of GCEs and CSEs) undoubtedly contributed to the process. Changes in examiners' assumptions and practices over the

same period may also have fuelled the rises. At the same time, however, there are two further sets of explanations which exam boards have historically defined as beyond their control and consequently scarcely consider at all.

First, the social composition of the pupil population has been changing. Better-educated parents have created higher expectations of performance, disadvantaged groups typically associated with lower performance have declined in size and young people themselves appear to have become increasingly motivated to take (and do well) in public exams. During the last decade the social significance of the 16+ examinations has also changed. In a situation where the majority of young people now stay on after their period of compulsory schooling has ended, one of the key function of public examinations may have changed—the securing of 5 or more high grade passes becomes simply a stepping stone to the next stage.

Second, the climate within which schools find themselves operating has been transformed. In particular, they have been encouraged in recent years to 'improve' their pupils' exam performances (Gray *et al.* 1999). Schools have consequently adopted a variety of strategies. They have, for example, entered more pupils for more examinations; identified and focused their efforts on 'borderline' candidates (mostly at the crucial grade D/grade C boundary); provided more support for pupils to revise and hone their exam techniques; and, on occasion, exploited the opportunities to select exam boards likely to give them the most favourable grades. In the process they have gnawed away at the edges of some of the exam boards' core assumptions. Even if, as seems likely, examiners are increasingly aware of some of these influences they appear to have few, if any, ways of taking them systematically into account.

Finally, there are a set of factors related to the examiners' own behaviour. Cresswell suggests that 'fluctuations in the standards represented by (examiners') qualitative judgements' may contribute to differences in exam outcomes. In other words, even when groups of examiners think in similar ways, they may implement their shared assumptions in somewhat different ways. Other factors may also be influential, however, even when exam boards are broadly aware of them. The changeover from one chief examiner in any particular subject to another, for example, could be a time of disjuncture as could other factors influencing the recruitment of particular cohorts of examiners. An unusually large number of appeals in the previous year or extensive

criticism of current standards may create a climate for the next year. And the fluctuating fortunes of different subjects as they 'compete' for pupils, with some rising and others falling in popularity, may also feed into the process. Differences between exam boards may also need to be brought into the reckoning. It is incumbent upon exam board officials to claim that such factors have been taken into account; unfortunately it is not always clear how they have done so.

Concluding Thoughts

The procedures currently adopted by exam boards to 'maintain standards' from one year to the next are fairly limited. They clearly take steps to use such evidence as they have readily to hand to inform their judgements. Indeed, within their own terms and as working hypotheses across short spans of years, such approaches can probably be said to work. During periods of rapid change, however, some of their assumptions are likely to break down. It needs to be acknowledged that the quest for truly 'objective' standards is illusory. Nonetheless, the challenge to existing practices thrown down by Cresswell's paper is clear. Should exam boards' efforts to 'maintain standards' continue to be almost entirely self-referencing? Or is it time to consider the introduction of a wider range of external evidence?

Lindsay Paterson

Introduction

Mike Cresswell's paper is a definitive demonstration that judgement is unavoidable when assessing standards. I have no disagreement with that. But I would take issue with the implications which his paper briefly indicates. At several points, he draws a sharp distinction between objectivity and judgement. Indeed he says that 'examination standards . . . cannot be made objective'.

My main point is that judgement is not just subjective. Judgement is socially constructed—as Cresswell acknowledges but does not develop—and as such can be the basis of social research. Moreover, that research into the social basis of judgement can be every bit as rigorous as the statistical analysis of examination results. In fact, it can be every bit as statistical as well.

Knowledge is Socially Constructed

So my first point is that judgement is not subjective. If it is not objective, then it is at least inter-subjective. Not being a philosopher, I am not really in a position to provide a full philosophical analysis of this debate. But I would make three points:

1 There are now many reasons to question the logical positivist claim that everything which is not observable (or derived from a priori reasoning) is mere opinion (Barnes 1974, Berger and Luckmann 1966, Hollis and Lukes 1982, Kuhn 1970, Lakatos 1974).

2 Being a sociologist rather than a philosopher- and being, moreover, a Scottish sociologist—I turn for guidance on this to the Scottish Enlightenment. A central theme of the Enlightenment's epistemology was to distinguish between judgement and mere opinion. For example, here is how Christopher Berry (1997) has recently characterised Adam Smith's view of the matter: 'through our imagination we are able to conceive what we would feel if we were in the situation of another' (p. 162). Smith called this 'sympathy'. As a result, 'individuals identify themselves with the "impartial spectator"' (p. 164)

3 And then we can find that theme appearing specifically in writing about statistical epistemology, for example in Ian Hacking's argument that the reason why we accept statistical method at all is social convention (Hacking 1965: 52, 226–7). An example is in the apparently arbitrary levels of statistical significance with which we are accustomed to operating, at least informally: Hacking's point is that these levels still matter, despite their arbitrariness, precisely because they have been so widely used in practice by natural and social scientists. And he cites C. S. Peirce arguing that, to make sense of statistical inference at all, we have to align our own interests with those of the whole community (Hacking 1990: 211).

The Social Uses of Credentials

The next point, then, concerns the implications of knowledge's being socially constructed. It follows that I am in complete agreement with Cresswell that norm-referencing is unavoidable. Attempts to establish criteria of assessment can work only locally, as it were, because the *selection* of criteria is socially conditioned. For example, our societies seem to be coming to the firm consensus that basic IT skills have to be acquired by everyone who could reasonably call themselves educated.

So that looks like a criterion. But the selection of that criterion is a matter of social judgement: there is nothing absolutely compelling about any particular level of IT skill. After all, just a few years ago, the criteria of what is required would have been quite different. The same point about the social selection of criteria can be made about other topics that might appear to be criteria—for example, learning to recite the Lord's Prayer or doing long division by hand.

Moreover, just as choosing criteria for assessment is socially conditioned, so too are the uses to which public examinations are put. Whatever educationalists might prefer in the way of criterion-referencing, we cannot get away from the social sorting role of public examinations—an aspect of what Randall Collins (1979) calls credentialism. It is not for reasons of obtuseness that employers or university selectors continue to insist on having ways of ranking applicants, and it is not unreasonable that they turn to public examinations as a first, crude measure of the rank order.

So, if norms and social sorting are inescapable, the questions for social research are things like this:

- on what grounds are the norms socially acceptable at any particular time?
- how do one set of norms lose social acceptability?
- how are new norms established?

Another way of putting this is to see the entire enterprise of public examinations as a vast piece of research in itself, which has as its primary social aim the discovery of the most effective way of allocating people to social roles, and of equipping them to move among social roles. So—translating the questions I have just asked into the language of research design—the problem is to establish the validity of this vast bit of research:

- do the examinations validly measure things which society judges to be worth measuring?
- does the examination system respond to changes in these social judgements about what is worth measuring?
- does society have ways of altering the examination system to incorporate new judgements?

An Illustration: Scotland

Establishing the validity of the research project of public examinations would itself require several large research projects. Let me illustrate the potential, however, by the instance of Scotland. Despite what Cresswell implicitly claims, his paper is not about Britain at all. It is about England, or maybe England and Wales. Indeed, I guess that a meeting such as that held at the British Academy in October 1998 could not have taken place in the same form in Scotland, because public examinations there have been subjected to far fewer criticisms and doubts than they have been in England.

That is partly a historical point. Scots quite like meritocratic selection, and have done so for at least a century. But it is also a more recent one. The crisis of confidence in standards has simply not affected Scottish debate to nearly the same extent as in England. Trying to use social research to understand why matters are different in Scotland can point to some features that would be needed in a programme of research on the social origins and development of standards. I cite this research, not to imply complacency about Scottish standards, but simply to illustrate how we might conduct research on the social acceptance of standards. My point is not to try to show that standards in Scotland really are higher or lower than in England, but to give examples of research that might be relevant to understanding the social basis of trust in current standards, whatever they may be.

To start with, let me note the evidence that there is no public sense of crisis. Repeated social surveys have shown that around two thirds to three quarters of the general population believe:

- that comprehensive secondary education is effective (as shown by analysis of the 1997 British Election Survey reported in Brown *et al.* (1999), by analysis of the British Social Attitudes Surveys reported by Arnott (1993) and by Paterson (1997), and by analysis of a poll conducted in 1996 by System 3 reported by Paterson (1997));
- that it teaches the basics well (as shown by analysis of the British Social Attitudes Surveys reported by Arnott (1993) and by Paterson (1997), and by analysis of the poll conducted in 1996 by System 3 reported by Paterson (1997));
- and that standards are not falling (as shown by a poll conducted by ICM for *The Scotsman* (1998));

One of the reasons for this belief is probably the research which has been carried out on the specific judgements made by individual examiners. The most notable example is the investigation by the Scottish Council for Research in Education in 1996 into the standards of the Higher Grade examinations (Devine *et al.* 1996). That was commissioned by a Conservative-controlled Scottish Office, and had as one of its effects the removal of the issue of examination standards from serious political debate. Over time, a more important source of research than the SCRE project has probably been the research which has been conducted by the former Scottish Examinations Board itself (now the Scottish Qualifications Authority). For example, since the late 1970s, it has monitored standards by statistical techniques devised by Alison Kelly (1976), essentially using factor analysis to study the relative difficulty of examinations in each subject in each year, and then allowing the conclusions to inform the setting of next year's papers. The Board and Authority have also retained a random sample of scripts in each subject at each level in each year to allow retrospective checking of standards (which was how the SCRE research was possible).

Most importantly of all, however, several programmes of research now offer explanations of the rise in examination attainment. This research is relevant in this context because it presents plausible explanations of these changes that do not depend on claiming a fall in assessment standards. Thus the research has helped to support the claim that the examinations are performing the purposes that society wants them to perform. This research addresses only one aspect of what would have to be addressed in a full programme of research into the social basis of examination standards: it addresses the question of validity—of whether the examinations continue to perform the selection functions that society wants of them. The research does not address, for example, the question of *how* standards are developed or change.

Four points emerge from that research on why examination attainment has risen. The first and most important is the second-generation effect of rising educational capital (Burnhill *et al.* 1990, Paterson 1992). Put simply: the children of better-educated parents do better, and there are a lot more better educated parents around now than there were thirty years ago.

Second, if we ask why today's parents started liking school two decades or so ago, Scottish research provides us with quite a specifically Scottish answer: comprehensive secondary schools (Benn and Chitty

1996, McPherson and Willms 1987). The evidence from research is that they did indeed partly (but only partly) fulfil one of the intentions of their founders, of tapping into educational talent in social groups which had previously been wholly excluded.

Third, comprehensive inclusion of the working class was made easier to achieve politically by the fact that these same disadvantaged social groups were declining in size (Paterson 1995). This purely demographic trend was intensified by social mobility—parents moving out of working-class occupations into middle-class ones, and their children then acquiring the educational characteristics of the middle class in general.

The fourth point is about how demand for education can be stimulated (Paterson 1992, Robertson 1992). Scottish research has shown that young people respond to what is on offer in higher education, and are motivated to attempt and pass examinations in order to take advantage of what is on offer. There is no 'fixed pool' of demand. And the same can be said in the post-compulsory stages of schooling (Paterson and Raffe 1995). Reforms to curriculum and assessment in the last four years of Scottish schooling have contributed to stimulating demand, by making these school years more enjoyable.

I have gone through this research-based explanation of rising attainment in order to indicate what research can do—what it can show about the validity of the public examinations as selection mechanisms. The research shows that there are good, sensible, and rigorously established reasons why young people are doing better, and so that panic about a crisis of standards is simply unnecessary. I should say also that all the research I have cited has been statistical, most of it using quite complex methods of analysis.

But even that research would not have been compelling if there had not been popular faith in the system. Education is regarded as a key Scottish civic institution (Paterson 1998; forthcoming); there is a long-standing trust in meritocratic selection as a fair and effective form of social organisation (Anderson 1995, Gray *et al.* 1983, Smith and Hamilton 1980); there is quite widespread trust in teachers' capacity to make the system and the selection mechanism work (as shown by the poll conducted by ICM for *The Scotsman* (1998)); and the system is trusted to reform itself to meet new demands (as shown, for example, by the near-unanimous view that decisions about Scottish educational policy should be taken in Scotland: see data from ICM reported in *The Scotsman* (1998)). The educational research I have rehearsed has probably helped to establish and maintain that trust.

The trust creates a social climate in which a principle of parsimony excludes an explanation in terms of declining standards. Provided there are adequate other explanations of rising attainment, the socially and politically inconvenient option of doubting standards is simply not needed. Whether that basis of trust will survive the creation of the Scottish parliament in 1999 remains to be seen. It could well be that a parliament with democratic legitimacy will erode the popular confidence which Scottish civic institutions have been able to command in the Union. (For a discussion of this, see Paterson (1998).) The point for us, though, is the methodological one that the same processes of social research would be able to measure a decline of trust as have been used to investigate its social basis hitherto.

Conclusion

Contrasting Scotland to England (and maybe Wales) points to the kind of research which could be done to find out what the necessary conditions are for the social acceptability of an examination system: that is, in relation to Cresswell's 'value judgements', establishing what their social origins are and what are their mechanisms of change. My methodological point, for academics and policy makers, is that doing this research is not at all an attempt to escape from the rigour of analysing examination statistics. If anything, the complexity of this task makes it rather more methodologically demanding than annually perusing the tables of crude examination pass rates.

A. H. (Chelly) Halsey

Mike Cresswell has written a splendid paper arguing that under the debate on *educational* standards lies a more technical and philosophical controversy about the reliability and validity of public *examination* standards. He points to research evidence on the fragility of long-term comparisons of examination standards, to the shifting perspectives of the guardians of those standards, the changing pressures from the interested parties and the rising intervention of the State. He argues that, in the end, it is all based, however rationally, on value judgement. I agree with him, but with modification.

First examinations, while often used, as he asserts, 'to provide information for future meritocratic educational and vocational selection

decisions' are also vital tests of competence, for example in the safe driving of cars or in surgery on human beings. The language of value judgement, other than universal accord that road accidents are to be avoided and human life preserved, is perhaps less relevant in such cases. The essential point of the examination for a driver or a doctor is to establish not who is better than whom but rather who can perform the job adequately. The task, in other words, is a classification of competence, not an ordinal ranking of competitors. From that point of view modern and future society may be deemed to need examinations for efficiency rather than for selection.

Second, again though I agree that debate over educational standards is underpinned by one over examination standards, I would make the sociological point that tests of competence to perform *any role* are part and parcel of social arrangements. They are by no means confined to formal schooling. For example monarchs and the inheritors of businesses reach their positions through 'ascription' but it should also be noted that 'achievement' is added to ascription as part of the expected preparation for the role, involving subjection to the exhausting ritual round of royal appearances in the one case and workshop experience, business school training etc. in the other. Examination may simply follow the rules of inheritance at death in either case, but the examination also involves abdication or bankruptcy with parliament or the receiver as examiners. Exploration of these complicated forms of ascription and achievement and their interactions would take us extensively beyond the limits of Cresswell's paper. It would, for example, raise questions of whether there should be examinations in capacity for parenthood and, if so, whether this preparatory education should be located in schools.

Third, however, I want to take up and elaborate Cresswell's insistence on the function of examinations for selection, whether meritocratic or otherwise, and illustrate it critically by reference to the expansion of higher education in the UK especially in 1992. Societies in general live by rewarding role players with money, honours, privileges etc and by punishing failure with imprisonment, demotion, removal of licence to practice a qualification, and many other kinds of disesteem. Against this background the universities have to be seen as loci of honours and privileges but in addition as emerging centres of the economy, the polity and society. Standards must therefore be maintained as to student entry, certification at exit, appointment and promotion of staff, and efficiency of learning and discovery.

Now all the conditions of supply and demand have been changing rapidly with respect to further and higher education over the past thirty years i.e. since the Robbins Report. The proportion of the relevant age group entering universities has multiplied ten fold since before the second world war and is now one third. On a still more long-range perspective there are now at least three times as many university teachers as there were students (20,000) at the beginning of the century. In 1985 11 per cent of the seventeen year old population obtained 3 or more A level passes. By 1995 this figure had doubled to 22%. Can it be concluded that standards of entry are declining or that there has been grade dilution (inflation)?

Cresswell, if I understand him, replies that the question cannot be answered and offers several technical reasons. By contrast the Dearing Committee (Dearing 1995) (with no discernible expertise of the kind amply demonstrated by Cresswell) concluded that there was no basis for the view that entry via A level to higher education had become significantly easier. Nevertheless it must be noted that in the same year the pass rate at A level rose again for the 16th year in succession, though the proportion of A's at A level levelled off. In 1998 the performance rate continued to climb but only just. From the standpoint of political and social function the reactions were highly if cynically predictable. The government congratulated itself for its reforms and commended teachers and candidates for their hard work, the teacher unions celebrated rising productivity and demanded increased pay, the opposition suspected falling standards in the shape of grade inflation.

Given the speed of expansion, conflict comes as no surprise. It expresses itself in diverse challenges to authority: for example over anonymity of refereeing papers submitted to journals for publication, or over judging applications for appointment or promotion, or over the validity of league tables, or over the award of ranks by central agencies to particular university departments. In short, the rapid replacement of tiny, consensual and élite universities by mass systems of higher education, whatever its great merits, leads to the decline of trust and to demand for greater openness in decision making.

But let us focus on matriculation—the conditions of entry into universities. The USA, leading an expansionist field, was first also in developing national standardised attainment tests. Each European country has a specific national educational qualification which forms the main basic requirement for entry to higher education. The qualification generally covers at least five subjects, some compulsory, and

usually including mathematics, the native language and one foreign modern language. England, Wales and Northern Ireland are unusual in limiting the number of subjects more narrowly and thus specialising earlier. At least five passes at GCE (usually taken at age 16) are required for degree level courses, of which two must be of Advanced level (usually taken at 18), although most candidates for entry attempt three A level subjects and already have at least 6 O-level passes. There is a passionate controversy over the special position of the A level examination in England, which guards entry to the university as does the abitur and the baccalauriat in Germany, France and elsewhere. Behind it lie the status and class battles for possession of educational property which have been intensified by the reform and expansion movements of the period since World War II. Special arrangements meanwhile exist for the growing body of mature students and those lacking 'traditional' qualifications.

In East and West Europe generally the state has increasingly controlled entry to higher education since Napoleonic times, either through defining examination content and standards or through varied means of student financial support or through special schemes of encouragement for particular social categories of student by positive discrimination or, more usually, by setting up barriers to entry. In the West some countries like Belgium, France or Germany used one uniform national examination. Sweden attempted the ranking of students by marks weighted according to the courses taken and work experience (which tacitly modifies age as a selective barrier). Positive discrimination in favour of candidates with working class backgrounds has been used in Poland and Czechoslovakia, as well as in Hungary, though examination performance has also been part of the entrance procedure. Entrance examinations have been widely used with higher requirements in medicine, science and law. Such procedures obtain not only in the highly prestigious institutions such as Oxford and Cambridge in England and the Grandes Ecoles in France, but also in the East European former Communist states where, at least a quarter of the places were reserved for working-class students. Even the lottery is not unknown. In the Netherlands the problem of excessive demand was overcome by its use. A lottery was operated in which an individual's chances were weighted by marks attained in the secondary school leaving examinations.

Nevertheless the automatic right of entry to the university which is the traditional privilege of those who obtain a baccalaureat or the abitur, still gives admission in France and Italy, though not to other

forms of higher education. The consequences are seen in high failure or dropout rates in the first two years of undergraduate study. We have here essentially a form of retrospective examination by actual performance 'on the job'. Even in England and Scotland this phenomenon has appeared since the expansion of the universities to include the former polytechnics in 1992. It is an inevitable consequence of the transformation to mass higher education. In other words it is possible to use the first one or two years of university study as a selective device in place of the traditional matriculation examinations of the upper secondary school. It is therefore not surprising that as late as 1994 there was fear of rioting in Paris and reports of long queues for admission in Bologna. Other countries, like Belgium or Spain, never granted the prerogatives of the abitur. In France, however, in spite of several university reforms, including the Loi Savary of 1984, the right of entry of a bachelier has never been modified. Of course, the highly selective Grandes Ecoles continue to cream off the best 15 per cent or so of the candidates. And the *numerus clausus* has been increasingly applied in France and Germany so that we can now describe the right as nominal. It does not guarantee a place in any particular faculty of any particular university.

In summary it appears that the evolution of matriculation and the admissions system in recent decades has been to move the point of selection upwards from the upper secondary school and its examinations to the admissions offices of the institutions of higher education. The traditional system was essentially controlled by teachers in universities. Control now is much more in the hands of politicians and budgetary administrators. Diversity is to be found at both the secondary and tertiary levels and the unique role of the baccalaureat, the abitur and their equivalents in other European countries as the *rite de passage* to university education, is no more.

Instead there have developed alternative modes of entry to a diverse set of post-compulsory educational and training institutions with the parallel development of vocational equivalents to A level, the baccalaureat and the abitur. In France there is a technical baccalaureat with 12 options as well as the traditional one with 8 sections and a proposed 30 option practical baccalaureat which, it is expected, will be taken in one form or another by 80 per cent of the secondary school leavers by the end of the century. Dr Cresswell's warnings against the comparability of examinations in different subjects are thus alarmingly compounded.

In most countries most students first enter full-time higher education aged between 18 and 21. At the end of the nineteen eighties the rate of full-time enrolment in this age group was more than ten per cent in over half of the OECD countries. However, older students are also admitted everywhere; in Germany a quota of places is reserved for them. In the Nordic countries, Austria, West Germany and Switzerland full time enrolment in 1990 was higher among persons aged 22 to 25 than among those aged 18 to 21. Reasons for starting first study in higher education later in life are many; some students pursue lower level further education full-time or enter employment; others may retake entry examinations and so increase the range of institutions which will accept them.

All in all then it appeared by the nineteen nineties that the articulation of the formal education system to the labour market in Europe was entering a new state of flux. It was not only that the macro-economic management associated with Keynes, Bretton Woods, and the left-wing planning governments of the nineteen fifties and sixties was collapsing. Nor was it only that the command economies of Eastern Europe were rapidly eroded at the end of the eighties. It was also that the sexual division of labour was now being comprehensively renegotiated, that the 'career' to which university admission had been traditionally a key with its life-long employment in a superior trade or profession, was disappearing. Part-time and temporary contracts were becoming normal, and not only for casual, unskilled and unschooled work but for professional and technical appointments. Europe, along with the rest of the advanced industrial world, was entering a profoundly different phase of the development of its economy and society and therefore of its educational arrangements.

Bibliography

Adams, J. (1912). *The Evolution of Educational Theory* (London, Macmillan).

AIE (1996). *Assessment in Education*, **3**(2).

Aldrich, R. (1995). *School and Society in Victorian Britain: Joseph Payne and the new world of education* (New York, Garland).

Aldrich, R. (1996). *Education for the Nation* (London, Cassell).

Aldrich, R. (1997). *The End of History and the Beginning of Education* (London, Institute of Education).

Aldrich, V. C. (1963). *Philosophy of Art* (Englewood Cliffs, Prentice-Hall).

Anderson, R. D. (1995). *Education and the Scottish People* (Oxford, Oxford University Press).

Arnold, Matthew (1863). *A French Eton*, reprinted in *The Complete Prose Works of Matthew Arnold vol. ii, Democratic Education* ed. R. H. Super (Ann Arbor, University of Michigan Press, 1962), pp 262–325.

Arnott, M. (1993). Thatcherism in Scotland: an Exploration of Educational Policy in the Secondary Sector (PhD Thesis, Strathclyde University).

Ayer, A. J. (1946). *Language Truth and Logic* Second edition (London; Penguin).

Baird, J. (1998). What's in a Name? Experiments with blind marking in A-level Examinations. *Educational Research*, **40**(2), 191–202.

Baird, J. and Jones, B. (1998). Statistical analyses of examination standards: better measures of the unquantifiable? (Associated Examining Board Research Report—RAC/780).

Bardell, G.; Fearnley, A. and Fowles, D. (1984). *The contribution of graded objectives schemes in Mathematics and French* (Manchester, Joint Matriculation Board).

Barnes, B. (1974). *Scientific Knowledge and Sociological Theory* (London, Routledge and Kegan Paul).

Bartholomew, D. J. and Knott, M. (1999). *Latent Variable Models and Factor Analysis* (2nd edition) (London, Arnold).

Bartholomew, D. J. (1996). *The Statistical Approach to Social Measurement* (San Diego, Academic Press).

Beardsley, M. C. (1981). *Aesthetics: Problems in the Philosophy of Criticism* (Indianapolis, Hackett).

Beaton, A. E. and Zwik, R. (1990). *Disentangling the NAEP 1985–86 reading anomaly.* (Princeton, Educational Testing Service).

Benn, C. and Chitty, C. (1996). *Thirty Years On* (London, David Fulton).

Berger, P. and Luckmann, T. (1966). *The Social Construction of Reality* (London, Penguin).

Berry, C. (1997). *Social Theory of the Scottish Enlightenment* (Edinburgh, Edinburgh University Press).

Best, D. (1985). *Feeling and Reason in the Arts* (London, Allen & Unwin).

Bierhoff, H. (1996). Laying the foundation of numeracy: a comparison of primary

school textbooks in Britain, Germany and Switzerland. *Teaching Mathematics and Its Applications*, **15**, 141–60.

Billington, R. (1988). *Living Philosophy: An Introduction to Moral Thought* (London, Routledge).

Bourdieu, P. (1989). *La Noblesse d'État: Grandes Écoles et Esprit de Corps* (Paris, Les Editions de Minuit).

Brock, M.G. and Curthoys, M.C. (1998). (eds.). *The History of the University of Oxford vol. vi, Nineteenth-Century Oxford, Part 1* (Oxford, Clarendon Press).

Brooks, G. (1997). Trends in standards of literacy in the United Kingdom, 1948–1996 (paper presented at the UK Reading Association conference, University of Manchester, July 1997, and at the British Educational Research Association conference, University of York, September 1997).

Brown, A., McCrone, D., Paterson, L. and Surridge, P. (1998). *The Scottish Electorate* (London, Macmillan).

Burnhill, P., Garner, C. and McPherson, A. (1990). Parental education, social class and entry to higher education, 1976–1986. *Journal of the Royal Statistical Society*, series A, **153**, 233–248.

Burstein, J., Kaplan, R., Wolff, S., and Chi, L. (1997). Using Lexical Semantic Techniques to Classify Free-Responses (Princeton N.J. Educational Testing Service Research Report available on ETSnet at http: //www.ets.org/research/siglex.html).

Christie, T. and Forrest, G. M. (1981). *Defining Public Examination Standards* (London, Schools Council/Macmillan).

Cipolla, C. M. (1969). *Literacy and Development in the West* (London, Penguin).

Clanchy, M. (1979). *From Memory to Written Record: England 1066–1307* (London, Edward Arnold).

Collins, R. (1979). *The Credential Society* (New York, Academic Press).

Committee of Council on Education (1863). *Report of the Committee of Council on Education 1862–63* (London).

Committee of Council on Education (1872). *Report of the Committee of Council on Education 1871–72* (London).

Committee of Council on Education (1873). *Report of the Committee of Council on Education 1872–73* (London).

Committee of Council on Education (1883). *Report of the Committee of Council on Education 1882–83* (London).

Copeland, J. (1993). *Artificial Intelligence: A Philosophical Introduction* (Oxford, Blackwell).

Cox, C. B. and Dyson, A. E. (1971). (eds.). *The Black Papers on Education* (London, Davis-Poynter).

Cresswell, M. J. (1987). Describing Examination Performance: grade criteria in public examinations. *Educational Studies*, **13**(3), 247–65.

Cresswell, M. J. (1990). Gender Effects in GCSE—Some Initial Analyses (Paper prepared for a Nuffield Seminar at University of London Institute of Education on 29 June 1990) (Unpublished Associated Examining Board Research Report—RAC/517).

Cresswell, M. J. (1994). Aggregation and Awarding methods for National Curriculum

Assessments in England and Wales: a comparison of approaches proposed for Key Stages 3 and 4. *Assessment in Education*, **1**(1), 45–61.

Cresswell, M. J. (1995). Technical and Educational Implications of using Public Examinations for Selection to Higher Education. In T. Kellaghan (ed.), *Admission to Higher Education: Issues and Practice* (Dublin, Educational Research Centre and Princeton, International Association for Educational Assessment).

Cresswell, M. J. (1996). Defining, Setting and Maintaining Standards in Curriculum Embedded Examinations: Judgemental and Statistical Approaches. In H. Goldstein and T. Lewis (eds.). *Assessment: Problems, Developments and Statistical Issues* (London, Wiley).

Cresswell, M. J. (1997a). *Examining Judgements: Theory and Practice of Awarding Public Examination Grades* (PhD thesis, University of London Institute of Education).

Cresswell, M. J. (1997b). Can Examination Grade Awarding be Objective and Fair at the Same Time? Another Shot at the Notion of Objective Standards (Unpublished Associated Examining Board Research Report—RAC/733).

Cresswell, M. J. and Houston, J. G. (1991). Assessment of the National Curriculum—some fundamental considerations. *Educational Review.* **43**, 63–78.

Cressy, D. (1980). *Literacy and the Social Order: reading and writing in Tudor and Stuart England* (Cambridge, Cambridge University Press).

Damasio, A. R. (1995). *Descartes Error: Emotion, Reason and the Human Brain* (London, Papermac).

Davis, E. (1993). *Schools and the State* (London, Social Market Foundation).

Dean, C. (1998). Standards are not parents' top priority. *Times Educational Supplement*, 9 October.

Dearing, R. (1995). *Review of the 16–19 qualifications* (London, Department of Education).

Dennett, D. (1993). *Consciousness Explained* (London, Penguin).

Department for Education and Employment (DfEE). (1997). *Excellence in Schools* (London, Stationery Office).

Department of Education and Science (1967). *Children and Their Primary Schools. A Report of the Central Advisory Council for Education (England).* ii (London, DES).

Devine, M., Hall, J., Mapp, J. and Musselbrook, K. (1996). *Maintaining Standards: Performance at Higher Grade in Biology, English, Geography and Mathematics* (Edinburgh, Scottish Council for Research in Education).

Devlin, K. (1997). *Goodbye Descartes: The End of Logic and the Search for a New Cosmology of the Mind* (New York, Wiley).

Dore, R. (1996). *The Diploma Disease.* 2nd edition (London, Institute of Education).

Dreyfus, H. L. (1992). *What computers still can't do: a critique of artificial reason.* (Cambridge Mass., MIT Press).

Eagleton, T. (1993). *Literary Theory: An Introduction* (Oxford, Blackwell).

Eiser, J. R. (1990). *Social Judgement* (Milton Keynes, Open University Press).

Elwood, J. and Comber, C. (1996). *Gender differences in examinations at 18+* (London, Institute of Education).

Firestone, W. A. (1998). A Tale of Two Tests: Tensions in Assessment Policy. *Assessment in Education*, **5**(2), 175–192.

Fletcher, S. (1980). *Feminists and Bureaucrats. A study in the development of girls' education in the nineteenth century* (Cambridge, Cambridge University Press).

Fogelin, R. J. (1967). *Evidence and Meaning: Studies in Analytic Philosophy* (London, Routledge).

Forrest, G. M. and Orr, L (1984). *Grade Characteristics in English and Physics* (Manchester;,Joint Matriculation Board).

Foxman, D., Ruddock, G. and McCallum, I. (1990). *APU mathematics monitoring 1984–88 (Phase 2)* (London, Schools Examination and Assessment Council).

Fremer, J. (1989). Testing Companies, Trends and Policy Issues: A current view from the testing industry. In B. R. Gifford (ed.), *Test Policy and the Politics of Opportunity Allocation: The Workplace and the Law* (Boston, Kluwer).

French, S., Slater, J. B., Vassiloglou, M. and Willmott, A. S. (1987). *Descriptive and Normative Techniques in Examination Assessment* (Oxford, UODLE).

Galton, M. (1998). Back to consulting the ORACLE. *Times Educational Supplement*, 3 July.

Gierl, M. J. and Rogers, W. J. (1996). Factor analysis of the Test Anxiety Inventory using Canadian high school students. *Educational and Psychological Measurement*, **56**, 315–324.

Goldstein, H. (1983). Measuring Changes in Educational Attainment Over Time: Problems and Possibilities. *Journal of Educational Measurement*, **20**, 369–78.

Goldstein, H. (1995). *Interpreting International Comparisons of Student Achievement* (Paris, UNESCO).

Goldstein, H. (1996a) (ed.). *Assessment in Education, 3,* 2. Special Issue: The IEA Studies.

Goldstein, H. (1996b). International Comparisons of Student Achievement. In Little and Wolf (1996).

Goldstein, H. (1999). Performance Indicators in Education. In D. Dorling and S. Simpson (eds.). *Statistics in Society* (London, Arnold).

Goldstein, H. and Cresswell, M. J. (1996). The comparability of different subjects in public examinations: a theoretical and practical critique. *Oxford Review of Education*, **22**(4), 435–42.

Goldstein, H. and Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology*, **42**, 139–167.

Good, F. J. and Cresswell, M. J. (1988a). *Grading the GCSE* (London, Secondary Examinations Council).

Good, F. J. and Cresswell, M. J. (1988b). *Differentiated Assessment: Grading and Related Issues* (London, Secondary Examinations Council).

Gould, S.J. (1984). *The Mismeasure of Man* (London, Penguin).

Gray, J., Hopkins, D., Reynolds, D., Wilcox, B., Farrell, S. and Jesson, D. (1999). *Improving Schools: Performance and Potential* (Milton Keynes, Open University Press).

Gray, J., McPherson, A. and Raffe, D. (1983). *Reconstructions of Secondary Education* (London, Routledge).

Green, A., Leney, T. and Wolf, A. (1997). *Convergences and Divergences in European Education and Training Systems* (Brussels, EC Directorate-General XXII (Education, Training and Youth)).

Green, A., Wolf, A. and Leney, T. (1999). *Convergence and Divergence in European Education and Training Systems* (London, Institute of Education).

Hacking, I. (1965). *The Logic of Statistical Inference* (Cambridge, Cambridge University Press).

Hacking, I. (1990). *The Taming of Chance* (Cambridge, Cambridge University Press).

Hambleton, R. K. and Zaal, J. N. (eds.) (1991). *Advances in Educational and Psychological Testing* (Boston, Kluwer).

Hargreaves, D. H. (1996). Teaching as a research-based profession: policies and prospects (Teacher Training Agency annual lecture).

Heath, A. F. and Clifford, P. (1990). Class inequalities in education in the twentieth century. *Journal of the Royal Statistical Society*, series A, **153**, 1–16. .

Holland, P. W. and Rubin, D. B. (1982). *Test Equating* (New York, Academic Press).

Hollis, M. and Lukes, S. (1982). (eds). *Rationality and Relativism* (Oxford, Blackwell).

Holmes, E. (1911). *What Is and What Might Be* (London, Constable).

Jencks, C. (1972). *Inequality: A Reassessment of the Effect of Family and Schooling in America* (New York, Basic Books).

Johnson, V. E. (1997). An alternative to the traditional GPA for evaluating student performances. *Statistical Science*, **12**, 251–278.

Kelly, A. (1976). A study of the comparability of external examinations in different subjects. *Research in Education*, **16**, 37–63.

Kilpatrick, J. and Johansson, B. (1994). Standardised Mathematics Testing in Sweden: The Legacy of Frits Wigforss. *Nordic Studies in Mathematics Education*, **1**, *6–30.*

Koretz, D., Broadfoot, P. and Wolf, A. (1998) (eds.). *Assessment in Education*, **5**(3) (Special Issue on Portfolios and Records of Achievement).

Kuhn, T. (1970). *The Structure of Scientific Revolutions*, second edition (Chicago, University of Chicago Press).

Lakatos, I. (1974). *Proofs and Refutations: the Logic of Mathematical Discovery* (Cambridge, Cambridge University Press).

Little, A. (1996) (ed.). *Assessment in Education*, *4*(1) (Special Issue: The Diploma Disease Twenty Years On).

Little, A., Wang Gang, and Wolf, A. (1995) (eds.). *Sino-British Perspectives on Educational Assessment* (London, ICRA, Institute of Education).

Little, A. and Wolf, A. (1996) (eds.). *Assessment in Transition: Learning, monitoring and selection in international perspective* (Oxford, Pergamon).

Long, H. A. (1985). Experience of the Scottish Examinations Board in developing a grade-related criteria system of awards (Paper presented at the 11th annual conference of the International Association for Educational Assessment held in Oxford, England).

Macaulay, Lord (1898). *Collected Works*, 12 vols. (London, Longmans Green).

Mackenzie, D. A. (1981). *Statistics in Britain 1865–1930. The Social Construction of Scientific Knowledge* (Edinburgh, Edinburgh University Press).

McKenzie, D. (1994). The irony of educational review. *New Zealand Annual Review of Education*, **4**, 247–59.

McLean, L. D. (1996). Large-Scale Assessment Programmes in Different Countries

and International Comparisons. In H. Goldstein and T. Lewis (eds.), *Assessment: Problems, Developments and Statistical Issues* (Chichester, Wiley).

McPherson, A. and Willms, J. D. (1987). Equalisation and improvement: some effects of comprehensive reorganisation in Scotland. *Sociology*, **21**, 509–39.

Madaus, G. and Raczek, A. (1996). Turning Point for Assessment: Reform Movements in the United States. In Little and Wolf (1996).

Menet, J. (1874). *A Letter to a Friend on the Standards of the New Code of the Education Department* (London, Rivingtons).

Morrison, H. G., Busch, J. C. and D'arcy, J. (1994). Setting reliable national curriculum standards: a guide to the Angoff procedure. *Assessment in Education*, **1**, 181–199.

Murphy, R. J. L. (1982). Sex differences in Objective Test performance. *British Journal of Educational Psychology*, **52**, 213–19.

Murphy, R. J. L., Burke, P., Cotton, T., Hancock, J., Partington, J., Robinson, C., Tolley, H., Wilmut, J. and Gower, R. (1996). *The Dynamics of GCSE Awarding: Report of a project conducted for the School Curriculum and Assessment Authority* (London, SCAA).

Newcastle Report (1861). *Report of the Commissioners appointed to inquire into the State of Popular Education in England*, PP 1861 XXI (ii) (London).

Newton, P. (1996). The reliability of marking of GCSE scripts: Mathematics and English. *British Educational Research Journal*, **22**, 405–20.

Newton, P. (1997a). Measuring comparability of standards between subjects: why our statistical techniques do not make the grade. *British Educational Research Journal*, **23**(4), 433–49.

Newton, P. (1997b). Examining Standards Over Time *Research Papers in Education*, **12**(3), 227–48.

Orr, L. and Forrest, G. M. (1984). *Investigation into the relationship between grades and assessment objectives in History and English examinations* (Manchester, Joint Matriculation Board).

Orr, L. and Nuttall, D. L. (1983). *Determining Standards in the Proposed Single System of Examinations at 16+* (London, Schools Council).

Paterson, L. (1992). The influence of opportunity on aspirations among prospective university entrants from Scottish schools, 1970–1988. *Statistics in Society, Journal of the Royal Statistical Society*, series A, **155**, 37–60.

Paterson, L. (1995). Social origins of under-achievement among school-leavers. In L. Dawtrey, J. Holland, M. Hammer and S. Sheldon (eds.), *Equality and Inequality in Education Policy* (Milton Keynes, Open University Press).

Paterson, L. (1997). Student achievement and educational change in Scotland, 1980–1995. *Scottish Educational Review*, **29**, 10–19.

Paterson, L. (1998). The Scottish parliament and Scottish civil society: which side will education be on? *Political Quarterly*, **69**, 224–33.

Paterson, L. (forthcoming). Scottish traditions in education. In H. Holmes (ed.), *Compendium of Scottish Ethnology, vol. 11* (Edinburgh, Scottish Ethnological Research Centre).

Paterson, L. and Raffe, D. (1995). Staying on in full-time education in Scotland. *Oxford Review of Education*, **21**, 3–23.

Payne, J. (1872). 'Why are the Results of our Primary Instruction so Unsatisfac-

tory?', *Transactions of the National Association for the Promotion of Social Science.*

Phillips, M. (1996). *All Must Have Prizes* (London, Little, Brown and Company).

Pirsig, R. M. (1974). *Zen and the Art of Motorcycle Maintenance: An Inquiry into Values* (London, Bodley Head).

Plewis, I. (1998). Inequalities, Targets and Zones. *New Economy*, **5**, 104–8.

Plewis, I. (1999). What's Worth Comparing in Education? In D. Dorling and S. Simpson (eds.). *Statistics in Society* (London, Arnold), 273–80.

Pole, D. (1961). *Conditions of Rational Inquiry: A Study in the Philosophy of Value* (London, Athlone).

Power, M. (1997). *The Audit Society: Rituals of Verification* (Oxford, Oxford University Press).

QCA (1998). *GCSE and GCE A/AS code of practice* (London, Qualifications and Curriculum Authority).

Reynolds, D., Creemers, B. P. M., Stringfield, S. and Teddlie, C. (1998). Climbing an educational mountain: conducting the International School Effectiveness Research Project. In G. Walford, *Doing research about education* (Lewes, Falmer Press).

Roach, J. P. C. (1971). *Public Examinations in England 1850–1900* (Cambridge, Cambridge University Press).

Robertson, C. (1992). Routes to higher education in Scotland. *Scottish Educational Review*, **24**: 3–16.

Rose, S. (1997). *Lifelines, Biology, Freedom, Determinism* (London, Penguin).

Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education*, **13**, 191–209.

Schools Council (1979). *Standards in Public Examinations: Problems and Possibilities*, Report from the Schools Council Forum on Comparability (London, Schools Council).

SEC (1984). *The development of Grade-related Criteria for the General Certificate of Secondary Education—a briefing paper for working parties* (London, Secondary Examinations Council).

SEC (1985). *Reports of the Grade-related Criteria Working Parties* (London, Secondary Examinations Council).

SEC (1986). Draft Grade Criteria. *SEC News Number 2* (London, Secondary Examinations Council).

SEC (1987). Grade Criteria—Progress Report. *SEC News Number 6* (London, Secondary Examinations Council).

Shavit, Y. and Blossfeld, H. P. (1993). *Persistent Inequality: Changing Educational Attainment in Thirteen Countries* (Boulder, Col., Westview Press).

Skolöverstyrelsen (1980). Quoted in J. Kilpatrick and B. Johansson (1994). Standardised Mathematics Testing in Sweden: The legacy of Frits Wigforss. *Nordic Studies in Mathematics Education*, **1**, 6–30. .

Smith, J. V. and Hamilton, D. (1980) (eds). *The Meritocratic Intellect* (Aberdeen, Aberdeen University Press).

Start, B. and Wells, K. (1972). *The trend of reading standards* (Slough, National Foundation for Educational Research).

Stedman, L. C. (1998). An Assessment of the Contemporary Debate over US

Achievement. In D. Ravitch (ed.), *Brookings Papers on Education Policy* (Washington DC, Brookings Institution Press), 53–119.

Stephens, W. B. (1987). *Education, Literacy and Society, 1830–70: the geography of diversity in provincial England* (Manchester, Manchester University Press).

Sutherland, G. (1973a). *Policy-Making in Elementary Education 1870–1895* (Oxford, Clarendon Press).

Sutherland, G. (1973b) (ed.). *Matthew Arnold on Education* (London, Penguin).

Sutherland, G. (1984). *Ability, Merit and Measurement. Mental testing and English education 1880–1940* (Oxford, Clarendon Press).

The Scotsman Education (1998). 30 September: 4–5.

Thom, D. (1986). The 1944 Education Act: the 'art of the possible. In Harold L. Smith (ed.), *War and Social Change: British Society in the Second World War* (Manchester, Manchester University Press), 101–28.

Vincent, D. (1989). *Literacy and Popular Culture: England 1750–1914* (Cambridge, Cambridge University Press).

Walden, G. (1996). *We Should Know Better: solving the educational crisis* (London, Fourth Estate).

Wang Binhua (1995). Comparing HSCE in the People's Republic of China and GCSE in England. In Little, Wolf and Wang Gang (1995).

Wang Gang (1995). The Development of Public Educational Examinations in China from 1980. in Little, Wolf and Wang Gang (1995).

Wiliam, D. (1996a). Meanings and Consequences in Standard Setting. *Assessment in Education*, **3**(3), 287–307.

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

Wilmut, J. and Rose, J. (1989). *The Modular TVEI Scheme in Somerset: its concept, delivery and administration* (Report to the Training Agency of the Department of Employment, London).

Wolf, A. (1995). Competence Based Assessment (Buckingham, Open University Press).

Wolf, A. and Steedman, H. (1998). Basic Competence in Mathematics: Swedish and English 16 year olds. *Comparative Education*, **34**, 3.

Wood, R. (1991). *Assessment and Testing: A survey of research* (Cambridge, Cambridge University Press).

Young, M. (1958). *The Rise of the Meritocracy 1870–2033* (London, Penguin).