

ALTERNATIVE CONCEPTIONS OF COMPARABILITY

Jo-Anne Baird

Abstract

Comparable examinations have to be at the same standard. But what do people mean by 'examination standard' and what kinds of comparability are expected? How is evidence to be gathered about these types of comparability and are all of these approaches valid? This chapter outlines different definitions of examination comparability used in England by academics and the expectations of the media and general public. The purposes to which assessment results are put are discussed, as the alternative conceptions of examination comparability are linked to the uses of the assessment results. Given that there are different approaches, some commentators have proposed that we should select a single definition of examination standards and stick to it, so that the system is clearer and false expectations are not raised about what the examination system can realistically deliver. Whether a particular definition of examination standards can be prioritised above others is considered, as well as the implications of so doing.

1 What does society mean by examination standards?

The word standard is used in a multiplicity of ways, leading to a great deal of confusion. As an example, Aldrich (2000) notes that the Department for Education and Employment White Paper *Excellence in Schools* (1997) has the following sub-heading in one section, 'Raising standards: our top priority'.

Aldrich points out that there is more than one way in which educational standards can be raised. To take a sporting analogy, high-jump standards can be improved by increasing the height of the bar that people have to jump over, or by raising the number of people who can jump over the bar. Educationally, raising standards can mean expecting more of students or expecting more students to be able to demonstrate performance at a given level. Herein lies the root of a problem.

Without being explicit about exactly what is meant by examination standards, many commentators are critical. Whilst some of the definitions of examination standards used are consistent with academic definitions (see later), some are ruled out by assessment specialists as too simplistic or not part of the standard-setting process. Let

us take a look at some media attacks on examination standards and consider the definition of standards being used and whether it is encompassed by a definition used in the assessment research literature.

1.1 The curriculum, questions or assessments are too easy

One way in which England's bar has been lowered, some claim, is by making the curriculum too easy. Professor Bernard Lamb, of Imperial College London, has been quoted as stating that the science and mathematics curriculum standards have been lowered so drastically that British students are a year behind foreign students when they start university (see Box 1). The standard-setting process begins with the definition of the curriculum that students will study and this is a matter of national importance.

Box 1

A new science GCSE, for instance, concentrates on topics such as genetically modified food and global warming rather than scientific theory. A level maths has been reformed to allow pupils to cover less challenging topics, while pupils taking a new maths GCSE can get an A grade without answering any of the hardest questions.

Julie Henry, Telegraph Education Correspondent (*Daily Telegraph*, 5 March 2006)

Likewise, there are complaints that some subjects are easier than others (Box 2). This also matters if our systems give equal credit for grades in different subjects, which the Department for Education and Skills (DfES) school performance tables and University and Colleges Admissions Service (UCAS) points systems do. For now, let us take these complaints as qualms about UK students learning the wrong things, although there is clearly another issue here regarding comparability between subjects.

Box 2

The rise in interest in psychology is a consequence of what people are perceiving, that maths and physics are harder and they can get better grades in psychology... It is easy to show that psychology is an easier A level than maths. It is incredibly worrying because maths and modern languages are subjects that the country needs.

John Dunford, General Secretary of the Secondary Heads Association (*Times Educational Supplement*, 14 August, 2003)

When it comes to setting cut scores – the process normally considered to be standard setting – it is too late to influence the curriculum design issues. By that stage, students may have been studying for their examinations for two years. However, involvement of senior examiners in the standard-setting process is a way in which the validity of the content of the examinations can be checked and the curriculum

altered for future years if things have gone awry. Likewise, inclusion of expert judgements in comparability studies can be used as a commentary on whether the curriculum is appropriate.

Other, similar, attacks on educational standards relate to the questions themselves. Even if the curriculum is appropriate, questions could be set on the easier aspects. Alternatively, the structure of the qualifications can be questioned, with some claiming that the type of assessment undermines the quality of students' achievements:

- coursework – 'GCSE coursework to be curtailed to stop internet cheats' (Taylor, 2006)
- multiple choice examinations – 'Pick A, B or C for a GCSE' (Mansell, 2006)
- modular examinations – 'Modular exams "damaging degree courses"' (Lightfoot, 2006).

Recently, the examination regulators defended the difficulty of A level questions (Figure 1).

All of the above can be viewed as a tug-of-war with, at one end of the rope, progressive education stakeholders who wish to modernise the curriculum, increasing diversity in the curriculum and widening participation in education. As such, the modernisers are trying to raise standards by increasing the number of people who can jump over the bar. The traditionalists are at the other end of the rope, trying to maintain the highly selective function of the qualifications and keep the bar at the same height or even raise it if too many people are jumping over it. All of this is a question of degree, as the curriculum *must* change to keep up with advances in knowledge. A level computing would not have been a feasible subject 50 years ago when A levels were introduced, but technology is now crucial to development of the economy.

Equally, the examinations fulfil a selective function for higher education and employment. Opinions are bound to differ regarding what should be taught in our education system – should we focus more upon scientific theory or move more towards evaluation of scientific evidence on, for example, genetically modified foods? Should students focus more upon speaking a second language or upon written grammar? Whilst crucially important questions for educational standards, they are beyond the scope of this chapter. Note, however, that when new aspects of the curriculum are introduced, there is less time for treatment of the old curriculum material. If we focus upon the traditional curriculum material then we are almost necessarily going to see a decline in skills in those areas over time because students also have to learn new things. Arguably, the content of the education curriculum receives too little serious attention and debate.

Figure 1 Regulatory body advertisement to congratulate A level students on their results (17 August 2006)

5. **Figure 2**

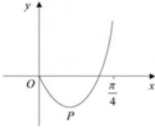


Figure 2 shows part of the curve with equation
 $y = (2x - 1)\tan 2x, \quad 0 \leq x < \frac{\pi}{4}$.


The curve has a minimum at the point P . The x -coordinate of P is k .

(a) Show that k satisfies the equation
 $4k + \sin 4k - 2 = 0$.

The iterative formula
 $x_{n+1} = \frac{1}{4}(2 - \sin 4x_n), \quad x_0 = 0.3,$
 is used to find an approximate value for k .


(b) Calculate the values of x_1, x_2, x_3 and x_4 , giving your answers to 4 decimal places.

(c) Show that $k = 0.277$, correct to 3 significant figures.




Congratulations to all A level students. You've tackled questions that would test the best of us.

7 Alexander Pope: Selected Poems
 "Effective satire is built on a foundation of irony." By comparing Pope's poetry with at least one other satirical work you have studied, discuss how far you agree with this claim.




2. European conflict and reconciliation, c. 1878-1980.
 "The desire for economic security was primarily responsible for determining European conflict and reconciliation between 1878-1980."
 How valid is this assessment of European conflict and reconciliation in the period 1878-1980?



SECTION E: COMPARATIVE PSYCHOLOGY





13 Discuss the role of social learning in the behaviour of non-human animals.



Questions taken from summer 2006 A level exam papers: Edexcel mathematics, OCR English literature, WJEC history, AQA psychology.

Today a quarter of a million students get their A level results. It takes most students two years of dedicated study to answer questions as tough as those above. So if you know anyone who's passed their A levels today, be sure to congratulate them. With the help and support of their teachers they have gained qualifications that are held in high regard and will stand them in good stead for university and employment.

If you would like to find out more about A levels, or would like more information about the questions above, visit www.qca.org.uk

The issues raised by challenges relating to the content of the curriculum and difficulty of the examination questions or assessment styles are subsumed by academic definitions of examination standards relating to qualitative judgements: criterion referencing and weak criterion referencing.

1.2 The pass marks are too low

There is a widespread assumption that grade boundaries are set at fixed proportions of the mark scale year on year, such as a pass being at 50%. The difficulty of the question papers varies from year to year, so to set the grade boundaries at the same mark every year would simply sanction examination difficulty varying between years. As the difficulty varies between years, the grade boundaries must compensate for this, to make it equally difficult between years to be awarded a particular grade. To put it bluntly, grade boundaries mean nothing in themselves. Nevertheless, there are occasional pleas for a simple system based upon fixed percentages of marks, with the proponents arguing that they are easier to interpret (Box 3).

Box 3

Teachers urge return to 'percentage' exams

Teenagers taking A levels and GCSEs should be given percentage marks instead of A, B and C grades, teachers urged today... Former PAT [Professional Association of Teachers] chairman Barry Matthews, who acts as an observer for the Government's exams watchdog, the Qualifications and Curriculum Authority, said he was concerned that pass marks changed every year... 'What I find difficult is that each year the examining boards (exam boards) adjust the actual pass mark. One year the pass mark for an A may be 70. The next year it could come down to 67,' he said. 'If my child got 68 this year and got a B and 68 next year and got an A, I would be concerned. The public would have more faith that exam standards were being maintained if the pass mark stayed the same every year.'... Wesley Paxton, on behalf of the union's education committee, put forward a motion demanding a return to 'numerical marks'... - Delegates passed the motion...

icWales.co.uk (27 July 2005)

<http://icwales.icnetwork.co.uk/0100news/0200wales/>

When an examination turns out to be far more difficult than anticipated, it would be perverse to penalise students by awarding them lower grades than they would have attained had they sat the examination in any other year. Equally, the examination papers may be designed to have particular grade boundaries at low marks. Until this year, GCSE mathematics examinations had three tiers of assessment, with the highest tier being designed for the most able students. As such, to get a grade C on that question paper, students would score only a few marks on very difficult questions.

Definitions of examination standards relating to pass marks are not addressed by any of the academic definitions in England because we have a system in which the pass marks are adjusted to compensate for changes in demand of the assessments. There

is a huge body of research literature on test equating and pre-testing. Creating assessments of equal difficulty year on year is an approach that could be feasible if pre-testing of the questions was conducted, but the high-stakes nature of UK public examinations and the high costs associated with pre-testing have mitigated against this approach. Nonetheless, this leaves the examination system open to attacks that can undermine public confidence if there is a lack of face validity of the boundary marks.

1.3 Too many students are getting the grades

Typically, complaints that too many students are being awarded the grades emphasise the selection purpose of the examinations, and positioning the bar at a particular height, to ensure that only a select few are able to jump over it. This is explicitly not the definition of examination standards in use by the current government (Box 4).

Box 4

We no longer have the quota system of 20 years ago, which condemned 30 per cent of pupils to failure each year, no matter their achievements. Today, hard work merits success, and high quality teaching is enabling every young person to grasp opportunities.

Jim Knight, Schools Standards Minister (17 August 2006)

<http://www.dfes.gov.uk/qualifications/news.cfm?page=0&id=105>

Box 5

The pass rate has gone up for the 21st year in a row and more pupils are getting A grades than ever before – about 20% of those taking the exams. ...

The former chief inspector of schools Chris Woodhead believes public exams like GCSEs and A levels are getting easier. 'The A level examination is not fulfilling the function that it should be, namely it is not identifying the most gifted students for top universities,' he told the Today programme on BBC Radio Four. 'When you look at the rate of increase and the fact that each year each new generation does do better then I don't think you are a cynic, you are just intelligently sceptical, if you raise questions about the nature of the examination. It can't all be down to better teaching, greater dedication, more intelligent students.'

Angela Harrison (BBC News Online, 14 August 2003)

<http://news.bbc.co.uk/1/hi/education/3150189.stm>

However, rising proportions of students passing the examinations in England has attracted criticism annually. The year 2003 was an interesting year. A level pass rates increased and were subjected to the usual criticisms of dumbing down (Box 5). Also in

2003, the proportion of Scottish students attaining the grades in Higher English reduced. Instead of this being interpreted as a raising of the bar, it was interpreted as a lowering of standards because the education system had failed to produce as many students who were capable of jumping over the bar (Box 6). Thus, viewed from one of these perspectives, there can always be claims that examination standards have gone down. Journalists looking to write about contentious issues can thereby always generate a story about standards in decline: a juicy critique of the government of the day. Whether too many students are being awarded the grades is clearly a legitimate question – the point is that changes in national examination statistics are not a good test if standards can be interpreted as having fallen whether the results go up or down.

Box 6

But there was concern last night when it emerged that a fall in the pass rate for Higher English is even worse than feared, with four out of 10 failing the exam. The true figure is understood to be around 60% this year. That is a further 2% lower than has previously been suggested and represents a 12% fall over the past two years.

Fiona Hyslop, the SNP's shadow education minister... called for smaller class sizes and a greater focus in schools on literacy to drive up standards in English, following allegations that some sitting the exam are 'barely literate'.

Jason Allardyce (*Scotland on Sunday*, 10 August 2003)

Concerns about the proportion of students passing the examinations fall into the statistical camp of assessment literature definitions of examination standards, of which two are considered below: cohort referencing and the 'catch-all' definition.

2 What do assessment specialists mean by examination standards?

Lack of treatment of the assessment field in England as an area where expertise is required, has encouraged the view that assessment users are experts in assessment. Teachers and examiners, for example, are subject matter experts and may or may not be assessment experts, as they may have little training in assessment. Often the training they receive is very focused upon the marking of a particular assessment. In the following paragraphs, the academic educational assessment literature is drawn upon to review the definitions of examination standards from assessment specialists. Let us start with the simplest first.

2.1 Cohort referencing

Under this definition, the same proportion of students are awarded the grades each year. Cresswell (1996) calls this the 'no-nonsense definition'. This is often referred to as 'norm-referencing', but in norm-referencing a population with known characteristics is used to contextualise performance on the current test. For example, in IQ testing a norm-group is used to develop the test such that it has known properties. The Wechsler Adult Intelligence Scale has a mean of 100 and a standard deviation of 15 for all of nine

age groups used in the norm-group. Care was also taken to ensure that the norm-group was stratified by sex, ethnic group, geographic area in the US and six occupational categories. When a classroom of students is tested for their IQ, we can compare their results on the same test as individuals, or as a group, with the population-norm because it is exactly the same test that has been used. In examination testing in England, the tests are different each year, for security reasons. There are, of course, other models that could be used: the content of the Wechsler Adult Intelligence Scale test is kept secure. Security matters because it is possible to study for a test in which it is possible to predict the content. Releasing the tests, or at least ensuring that teachers and students are aware of their likely content, is therefore important to educational assessment in England. Norm-referencing gives information that can be used to compare an individual, or group, with a larger population.

William (1996a) points out that what in England is often called norm-referencing is more correctly called cohort referencing. In cohort referencing, examination standards are cohort specific – they do not tell us how the examination standards compare with last year, between subjects, between different examining boards and so on. This only matters to the extent that the purposes to which examination results are put depends upon being able to draw inferences about these different types of comparability. If examination results were used in a subject-specific manner, with those making inferences from them being knowledgeable about the different syllabuses offered by examining boards and only selecting between students in the same year group, then cohort referencing might well be a suitable system.

Unfortunately, this is not the case in England. Applicants for jobs and university places are drawn from different cohorts, and it would be impossible for employers or admissions tutors to be familiar with the content and demands of the 251 A level and 301 GCSE syllabuses currently on offer. Further, on the assumption that comparability exists between syllabuses, subjects, boards and qualifications, examination results are converted into points which are used in school performance tables and for entrance to universities.

Despite these weaknesses, Goldstein (1986) advocated such a system. So what are its benefits? It is a relatively simple system to deliver and for the public to understand. Moreover, there are those who argue, as this chapter demonstrates, that the present system is not only complex, but fails to deliver the kinds of comparability that it claims to. So why keep up this charade? Why not adopt a simpler system such as cohort referencing with all its known inadequacies?

In theory, this definition could be adopted in England, but in practice, there would be many powerful critics who would not tolerate it. The Working Group on 14–19 Reform (2004) rejected it as an option for standard-setting for the proposed diploma system (para 183):

However, grading of diplomas should not be norm-referenced over time. If an increasing proportion of young people meet the established criteria for higher grades then the proportion achieving those grades should be allowed to rise.

Teachers and schools would object to the lack of comparability between examining boards, as, under this system, an easier syllabus in one examining board would have the same distribution of examination results as a harder syllabus in another. If the harder syllabus also happened to attract more able candidates, then the disparity between what would have to be achieved to attain the grades would be compounded. Progression to further study or into employment may depend upon the knowledge gained from a course, and if that varies wildly, assumptions about what is learned from a course are not possible. Therefore, lecturers in further education and higher education would have complaints. Evaluation of the education system over time would not be feasible – a fixed proportion of candidates would pass the examinations every year, despite government policies on *Excellence in Schools* (DfEE, 1997), national numeracy and literacy hours (DfES, 2002) or specialist schools (DfES, 2006). Different measures of the impact of government expenditure would be needed. To work the system, cynical students and teachers would select subjects and syllabuses that had a low-ability entry. Thus, dumbing down would be an educational consequence. Simple cohort referencing was used loosely in the early A level and O level examinations, but disenchantment with the lack of information from such a system moved examination boards to change the system (see Chapter 2).

Cohort referencing is the simplest form of statistical approach to the definition of examination standards possible. As no inferences can be drawn from it about candidates' performances in the examinations, very little information is gleaned about examination comparability at all. Nonetheless, it could be argued that this would be a clear approach to the setting of examination standards that has no pretence about delivering more than can be delivered by any system. Educational reform would be difficult under this system, as new qualifications would struggle for recognition. Let us turn next to a more complex statistical definition of examination standards that attempts to address some of the problems of this simplistic method.

2.2 The catch-all definition

Under this definition, we would say that two examinations were of comparable standards if students with the same characteristics were awarded the same grades on average, no matter which examination they entered (Cresswell, 1996). By 'the same characteristics', it is intended that all characteristics that have a legitimate relationship with examination results are controlled for when comparing the outcomes of the examinations. We would take into account how able the students were who entered for each examination, as well as the quality of the teaching, motivation of the students, number of hours spent studying and so on. Chapter 10 discusses studies that have used this definition as the basis of their research into examination comparability. In practice, there are serious problems about measuring *all* relevant factors, but even if we set them aside, there are considerable theoretical difficulties with this approach.

Put simply, what does it *mean* to be equally prepared for different examinations? I had a brilliant mathematics teacher who was inspirational, but my biology textbook was engagingly modern, well-structured and easy to learn. How do we measure the

quality of teaching in mathematics and biology and how do we measure the quality of textbooks? But let us imagine that these problems were solved and we had all of the measures we needed to hand. All that would remain would be to conduct the statistical analysis that controlled for these factors so that we could investigate examination comparability. On constructing this model, if I find that for each hour of studying, candidates do better in mathematics than in physical education, do I conclude that mathematics is too easy? Using this catch-all definition, the answer is yes, as it has to be assumed that the relationship between these controls and the examinations being compared is the same for each examination. As soon as the relationship is allowed to differ, we cannot disentangle the examination difficulty from the supposed control for candidate preparedness. Naturally, life is not like this and these controls do vary between examinations. Girls do better than boys in GCSE chemistry, but worse in GCSE physics. Putting gender into an analysis comparing those examinations would therefore be highly problematical. This argument is more fully explored in Baird & Jones (1998).

Surely, though, there are some more-similar examinations where this definition would be useful? But even this is highly problematical. Examining boards have routinely carried out analyses using rudimentary approaches to this definition to compare different options within examinations – typically coursework with optional written papers (e.g. Massey & Baird, 2001). Having controlled for candidates' performances on the other question papers, candidates who take the coursework route often tend to do better. This could mean that coursework is too easy, but there is more to it than that.

In some examinations, candidates withdraw from the coursework option very late and enter for the optional written examination instead. This pattern, and examiners' experiences of these students, leads us to the question of whether lazy students, who have not completed their coursework, switch to the written paper in some of the qualifications. Completing the coursework may not only involve more effort, but it is possible that students learn more through their experience of that kind of assessment than an examination paper. What the relationship between the control variables and the examination results *should be* is not an empirical or theory-driven question, it is a value judgement.

There are several definitions that can be seen as sub-sets of the catch-all, but of course they are inadequate theoretically precisely because they do not attempt to measure all possible legitimate influences upon the examination results. What all of these approaches have in common is a rationale that attempts to control for the entry characteristics of candidates sitting the examinations. Essentially, the experimental design involves analysing the examination results, having controlled for differences between the candidates taking the two examinations.

Chapter 9 investigates the problems with assuming that a group of candidates should be awarded the same results in two examinations they entered. Another variant is assuming that similar schools ('common centres') should have similar results and the problems with that are outlined by Cresswell (1996). Yet another possibility is using a

reference test to control for differences in entry between two examinations, and this is discussed in Chapter 8. Robert Coe expands upon these possible models further in his commentary following this chapter.

A famous example of the inadequacy of a purely statistical definition arose in the introduction of the Curriculum 2000 A level examinations. Outcomes were predicted statistically, on the basis of candidates' prior attainment (mean GCSE scores). An assumption was made, not unreasonably one might think, that the relationship between prior attainment and A level grade would be similar for the new examinations compared with the same subject in the old-style A levels: a value-added definition. Actually, many assumptions have to be borne out for such a projection to be adequate. Equal teaching in both years, equal motivation of students, equivalent quality of textbooks and so on are necessary for this to be a reasonable assumption. In the first year of a new syllabus and examination structure, these assumptions simply do not hold.

The structure of the examinations had changed in the syllabus revision with the introduction of modular examinations (as proposed by Dearing, 1996), the main structural change being that candidates could certificate with an AS examination one year into the course. Students were encouraged to sit four AS level subjects in the first year and focus upon three subjects in their second year. The AS results were aggregated with the second year (A2) results to compose the new A levels. Certainly, the statistical predictions had operated well in the first year, 2001, with senior examiners' qualitative judgements of candidates' performances largely corresponding with the statistical information. Suffice it to say that setting of the AS standards was not generally problematical. When it came to setting the A2 examination standards, the boundary marks that would have been required to produce the statistical predictions were unacceptable to the senior examiners and were, frankly, ludicrous.

To illustrate the problem, in AQA A level French, 76% of students who sat the first AS examinations in 2001 went on to take the first A level examinations in 2002 (Table 1). Results for all AS students in 2001 had been similar to those for all A level students in the previous year, after controlling for prior attainment. Statistical predictions for A level results in 2002 were higher than A level results in 2000 (the last year in which only the old-style A levels were available), as the prior attainment scores (mean GCSE) for the cohort entered for the 2002 examinations were better. To achieve these statistical predictions, the boundary marks would have had to be very low at grade A and very high at grade E. On the oral examination, the difference between the highest and the lowest grade boundaries would have been 8 marks out of a total maximum score of 70. Either the assessments were very poorly designed – and that would have applied across all subjects – or there was something amiss with the assumptions underlying the statistical predictions.

From the actual 2002 A level results (Table 1), it is evident that it was concluded that there was something wrong with the statistical predictions. Investigations showed that there was an enormous disparity between students who dropped a subject at AS

Table 1 AQA French: statistical predictions and actual results

			Grade A	Grade E	Total number of candidates entered
2000	A level	Actual results	23.8%	88.7%	5,321
2001	AS	Actual results – all candidates	25.8%	94.1%	5,100
		Actual results – candidates who went on to study at A level	44.1%	99.8%	3,856
2002	A level	Statistical predictions	24.8%	92.4%	4,019
		Actual results	28.5%	97.2%	3,246

and those who continued to take it to A level. In this case, almost all AS candidates who continued to A2 study passed the AS examination (99.8%) and 44.1% were awarded a grade A at AS level. As a grade A at AS level accumulated enough points for a grade E at A level, 44.1% of A level students had passed the A level examination before they even sat an A2 examination and many more did not need to pass the A2 examinations to pass the A level – they only needed to score a few marks on each question paper (Pinot de Moira, 2002). One interpretation of this information is that the new assessments did not fit the new A level structure and that candidates did not deserve better grades than candidates with the same prior attainment had been awarded in the past. But a value-added definition is a weak version of the catch-all definition and more information was available.

A lower proportion of 18-year-old students had gone on to sit A levels in 2002 than in the previous year (Table 2). The introduction of AS certificates had a dramatic impact upon students' routes through the education system, with those who had been awarded better grades than they expected on the basis of their prior attainment being more likely to continue to the second year of study than those who had been awarded worse grades than they expected. Students were most likely to drop the subject in which they achieved their worst grade (Baird *et al.*, 2003).

Table 2 Proportion of 17-year-olds taking AS examinations and 18-year-olds taking A level examinations in England

	2001	2002
AS	40.0%	48.9%
A level	36.4%	35.7%

Source: calculated from Department for Education and Skills and National Office of Statistics figures

So another interpretation of the statistical information is that the change in examination structure had given students sufficient feedback about their strengths and weaknesses to make better choices about what to continue studying, thereby weeding out students who were not likely to pass the final qualification. Statistically speaking, putting the information we have about students' performances in AS

French into the statistical model would get us closer to the catch-all definition and we would not necessarily expect students' A level results in 2002 to be similar to those in previous years.

But hang on. The statistical value-added model that the predictions were based upon was created from old-style A level results, and the AS was not a feature of the previous A levels, so there was no such possibility. This highlights one of the problems with the catch-all definition: the world changes. If this entails changes in the relationship between the factors that are being used to control for entry between the examinations and the outcomes of those examinations, then the catch-all definition leaves a gap, as it is not possible empirically to know what those new relationships should be. Again, this has to be determined by value judgement.

Choice of factors to put into these statistical models is also a matter of values. If this were left to empiricism, a host of factors are related to examination results that we would think nonsensical to use to predict how students should be awarded grades. For example, anger (Lane *et al.*, 2005), physical attractiveness (Zahr, 1985), comfort of clothing (Bell *et al.*, 2005) and unattractiveness of first name (Erwin, 1999) have all been found to have predictive relationships with academic achievement. Other factors that are more traditionally reported with examination results, such as type of school, ethnic group, age, socioeconomic status and gender may seem more sensible, but we cannot fool ourselves that they are innocuous. Our choice of factors to put into these models represents our values about legitimate relationships with examination results.

By using these factors as controls, we interpret them as legitimate, but an alternative interpretation would be that they are biases in our examination system, and other analyses of the results, by gender or ethnic group for example, do indeed draw these conclusions (e.g. Gillbourn & Youdell, 2000). Now, the assessment specialist cannot control or even very much influence the relationship between these factors and examination results, but which factors are selected matters because students are not randomly allocated to examinations – there is choice. So, if an examination happens to have good results and is sat by a disproportionately higher number of females, then we may conclude that the examination is appropriately graded, but that females do better in examinations than do males. Other researchers may draw the conclusion that the examinations are biased in favour of females. Disentangling the examination standard from the features of the candidates who take the examination is impossible (Baird *et al.*, 2000).

Statistical literature on the best approach to setting up models of data abounds with debate on strategies for selection of factors to include, and criteria for so doing. Raudenbush (1994) argues that researchers should have theoretical reasons for the inclusion or exclusion of variables from models. Fishing expeditions, where researchers include anything that happens to have a significant effect on the model have, he notes, been found to overestimate the values of the coefficients in the models and underestimate the standard errors when cross-validation studies have been conducted. As previously discussed, theoretical reasons for relationships

between a disparate range of variables and examination results can be found in the literature.

Disappointingly, theory does not provide the whole answer that we seek, and Raudenbush alludes to this, when he says that there is ‘a decision about how to select variables for an analysis’ and his paper ‘has little to say about this important decision. Rather, the variables are viewed as given and a general approach to their analysis is prescribed.’ Choosing the variables for analysis is an experimental design question, involving selecting variables that not only have an empirical relationship with the examination results, but that we consider should be used as controls. After all, this experiment has social justice issues running through the design of it. A different dependent variable may help to illustrate the issues.

Imagine that we wished to use the catch-all definition to investigate fair pay, comparing two occupations: nursing and plumbing. Naturally, we would select control variables that, theoretically speaking, we would expect to have an empirical relationship with pay, such as number of years’ experience, amount of time spent training, work-related benefits (e.g. pension, sick leave), ethnic group and gender. All of these variables have significant effects in the model I create. For example, women get lower wages, as do people with a non-white ethnic background. I conclude that, having controlled for these variables, plumbers and nurses are paid equitably. Putting variables in the model for the purposes of exploring empirical relationships is a separate issue from using variables as controls for differences in the types of groups associated with the dependent variable.

Many would object to my model on the grounds that it is not fair that women and ethnic minority groups are paid less, and that this is not an explanation or a good control for differences in pay between nurses and plumbers. Indeed, setting up models like this and interpreting them in this way serves to compound existing inequalities if people accept these discrepancies as empirical and therefore legitimate. This example serves to highlight that the question of what it is legitimate to control for is not an empirical question, it is a value judgement. Value judgements change over time and depend upon individuals’ principles – an uncomfortable reality for researchers to accept.

Challenges to the adoption of the catch-all as the sole definition of examination standards arise because no methodology would be able to encompass all possible factors that could be included in a model. Even stronger challenges arise because this definition fails to consider the content of students’ performances.

2.3 Criterion referencing

With all of these problems with statistical approaches to the definition of examination standards, the answer to some is obvious – document the criteria that we expect of students’ performances and allow subject matter experts to judge the quality of students’ work against them. Sir Keith Joseph favoured this system when GCSEs were introduced in the 1980s (see Chapter 2) believing that it would measure ‘more

absolute standards' (Joseph, 1984). This definition is attractive, as the criteria would be available to educational stakeholders as an illustration and explanation of candidates' grading. Sir Mike Tomlinson's view of examination standards is similar (although he does allow that statistical information is also necessary for setting standards):

The basis for any system of assessment intended to judge students' achievement using a fixed standard should in principle lie in a comparison of individual students' work against that standard.

Tomlinson (2002, p. 25)

Also underlying this approach is the idea that it should be entirely feasible for examiners, indeed teachers, to make explicit the criteria for attaining particular grades. Teachers must be able to check students' learning by means of some form of assessment and explain why they have failed to make the grade. Students' performances would exemplify these written performance standards at the appropriate grades. Any notion that this is problematical would appear to undermine teachers' expert status, but experts' status in other areas has long been under threat – the literature shows us that doctors are not highly accurate at diagnosis, for example (Dowie & Elstein, 1988). Expert status is now more openly questioned in many areas of life. Like the simple cohort referencing system, this definition seems to have transparency in its favour. Many vocational qualifications have been developed with criterion referencing underpinning them. Wolf (2002) writes of the development of National Vocational Qualifications (NVQs) that the National Council for Vocational Qualifications' theory

... was that standards could be so clear and all-inclusive that anyone, in any factory, office or playgroup, would be able to use them to assess and measure performance accurately. As reality stubbornly failed to fall in with NCVQ's vision of perfect clarity, the level of detail required by the Council and the complexity of standards layout increased. It became more and more desirable for industries to acquire... the services of an experienced, all-purpose standards writer from the government's approved list.

Wolf (2002, p. 74)

In practice, delivering a transparent criterion referencing system does not easily translate into the kind of educational standards people expect to ensue. Criterion referenced examinations were introduced in New Zealand in 2004, following a significant teacher training programme to ensure that the standards were widely understood (Gilmore, 2002). The pass rate in the scholarship examinations dropped to half that of the previous year and there was an outcry over the overall pass rate, as well as variability between subjects. Approximately three-quarters of students sitting the Maori and Chinese examinations passed but, at the other end of the spectrum, the pass rate for physical education was 0% (Kingdon, 2005). The New Zealand Qualifications Authority's internal review (Martin, 2005) commented as follows:

This experience highlighted two problems. These were the level at which the standard was pitched and how realistic that level was. Setting an examination paper requires clarity about what is expected of students at a certain age (or a certain level of learning).

The issue is whether the standards address a certain level that the best students can realistically reach, or whether they are aspirational and aim at a level that the ideal student ought to reach. It is not clear which approach chief examiners and markers put in practice for the 2004 Scholarship, and whether each subject had the same notion of the standard.

Martin (2005, p. 11, paragraph 61)

Curiously, in criterion referencing, students are generally awarded the grade according with their worst performance, as all of the criteria have to be met to be awarded a grade (Forrest & Shoesmith, 1985). Rather than celebrating students' achievements, criterion referencing accredits students at the level of their weakest skill. Attempts to introduce criterion referencing in GCSEs failed because some students who deserved particular grades according with senior examiners' judgements did not meet the criteria (Cresswell, 1996). Compensation for weak performance in one criterion may be made by good performance on another criterion. Wilmot & Rose (1989) give the example of students who were good at seeking out information, but poor at communication and vice versa.

So far, only the problems associated with the criterion referencing system itself have been considered. The pattern of results in New Zealand is not surprising in relation to the educational assessment research literature and there are reasons to question the capacity of any judge to carry out the criterion referenced judgements in a way that is fair to candidates. This is discussed further in relation to the next possible definition of examination standards. For now, suffice it to say that all of the problems applying to human judgements of standards apply to criterion referencing, the main problem being a lack of adaptation of what is required from students depending upon the difficulty of the task being set.

2.4 Weak criterion referencing

Every year, over five hundred committees of eight senior examiners (on average) are convened to make judgements of students' performances on GCSE and A level examinations in England (Baird & Dhillon, 2005). The judgements being made are expected to take into account the difficulty of the examination. If setting examinations of equivalent difficulty was not problematical, these committees would not be required to meet, as the same grade boundary marks could be applied to the examinations every year. Under this definition, students' performances are said to be equivalent if they are of equal merit, in the judgement of senior examiners, after they have taken into account any changes in demand of the assessment. In practice, this is not the only definition of examination standards being adopted by these committees (Baird *et al.*, 2000), but there was a time when statistical information played a much less prominent part in the standard-setting process and it is useful to look at a case from that era, as it illustrates the problems that can arise under this definition. Research evidence on examiner judgements relevant to this definition of examination standards will then be outlined.

A case of weak criterion referencing in practice

In 1991, the Associated Examining Board's (AEB's) A level English examination (syllabus 0652) changed. One of the three question papers was marked out of 80 in

1991, rather than the 100 marks available in the previous year. However, the marking criteria were unchanged and the senior examiners 'could discern no way in which the marking schemes could have adversely or unfairly affected the results' (Day, 1992). The result of the senior examiners' grading judgements was to drastically reduce the proportion of candidates being awarded the top grades, but to increase the proportion of candidates passing (Table 3). So, if the assessment was of equivalent difficulty to the previous year, the candidates or the teaching must have been very different.

There was no evidence from the statistical analysis of the entry for the examination that the explanation was to be found there. The grade boundary marks set were very similar to those set in 1990 (after accounting for changes to the maximum mark) at grades A and B, with large reductions in the outcomes at those grades. If a similar grade boundary mark had been set at grade E, it would have produced 60.1% of candidates passing, but the grade E boundary mark was set at a lower proportion of the maximum mark than in 1990. This demonstrates that using the same boundary marks year on year has unpredictable effects. The scale upon which the boundary marks are being set changes, so it becomes like using a measuring tape made of elastic. The outcomes may or may not be similar to the previous year.

Table 3 AEB A level English (syllabus 0652) results

	Grade A	Grade B	Grade E	Total number of candidates entered
1990	5.7%	15.2%	60.1%	4,401
1991	0.7%	4.8%	71.9%	3,680
1991-1990	-5.0%	-10.4%	+11.8%	-721

Three schools protested to the Independent Appeals Authority for Schools Examinations (IAASE), who instructed the examining board to reconvene the awarding meeting, stating that:

The Authority was not satisfied with the overridingly judgemental nature of the award and it found that, statistically, the final grades awarded were out of line both with the marks produced by a team of experienced assistant examiners and with the awards in previous years on that syllabus. The Board had not given proper weight to the statistical evidence.

IAASE (1991)

The reconvened meeting included senior examiners from another AEB A level English syllabus (660). The entire panel was given a presentation on the statistical information and then conducted a thorough, independent review of the candidates' work, before producing new grade boundary marks. The new recommendations involved a small reduction in the grade A boundary marks, giving 1.1% of candidates a grade A. None of the upgraded candidates came from the schools who had appealed. The grade B and grade E boundary marks were not revised at all. The examining board documented the rationale for the grade boundary marks, indicating

how candidates had performed at each of the key grade boundaries. AEB's procedures in those days gave little authority to the board officers to challenge the grading judged by the senior examiners. The schools took their appeal back to IAASE, and AEB was heavily criticised:

The Authority, like the School and the LEA, was not satisfied with this response... It did not accept the Board's distinction between 'procedures' and the 'decisions' of its awarders: nothing in the report of the reconvened awarding meeting had altered the Authority finding; the standard of judgement which was applied by the awarders had been inconsistent with the standard applied in all other cases.

IAASE (1991)

This case resulted in a change to AEB's procedures, with statistics being given a more prominent role and awarding committees making recommendations to the chief executive, giving him final authority and accountability for the grading of the examinations. IAASE was clear that statistical information should have played a part in the process, showing that our educational institutions and structures are prepared to point to statistical definitions to challenge examiners' judgements when they believe the outcomes to be unjust. What this case and the New Zealand examination results have in common is the lack of reference to statistical information, and there is systematic research evidence, outlined below, showing that a reliance upon qualitative judgements will produce large swings in examination outcomes.

Research on examiner judgements

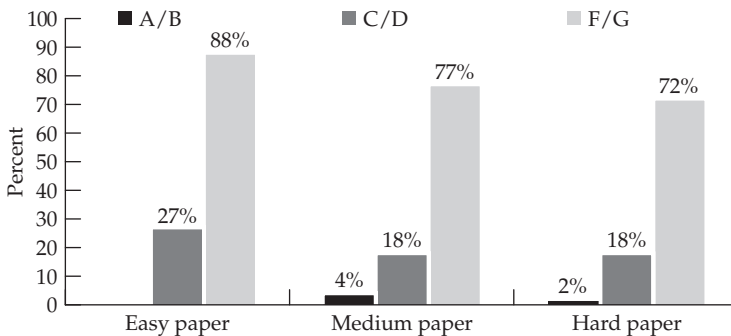
Cresswell (1997, 2000) analysed 108 grading decisions made in the 1994 examinations, comparing the boundary marks set by the examiners with those that would have been set to produce statistically equivalent outcomes. With random fluctuations in the sample of students taking examinations in any one year, it might be expected that there would be some changes in outcome and that they would reflect a normal distribution: most changes in outcomes would be small and there would be few extreme changes. Cresswell found exactly the opposite. He found few *small* changes: most were large swings in outcome compared with the previous year. These large swings were not explained by changes in the demographic nature of the candidates entered for the examinations, and they were not part of an ongoing trend.

Fortunately, the matter was not explained simply by the examiners having chosen the same boundary marks every year. There was clear evidence that examiners had responded to changes in difficulty of the examinations, with 77% of the boundary marks moving in the direction predicted by the statistical evidence. In fact, examiners tended to produce boundary marks that went halfway between the previous year's boundary marks and where the statistical information suggested the boundary marks should lie.

Furthermore, there is abundant evidence that examiners are not good at discerning the difficulty of questions (e.g. Impara & Plake, 1998) and question papers. Good & Cresswell (1988) investigated examiners' ability to set grade boundaries on tests that had specifically been designed to be easy, medium and hard and which were sat by

the *same* group of candidates. When candidates sat an easy paper, their performances were judged to be worthy of higher grades than when they sat the harder papers. Figure 2 shows the findings for the physics papers, but the same effects were found in French and history. So the reason that there is such variability in outcomes when a weak criterion referencing definition is adopted is that examiners cannot adequately compensate in their judgements of candidates' work for the demands of the question papers.

Figure 2 Grading of physics papers



Note: grade A was not available on the easy paper

Source: Good & Cresswell (1988)

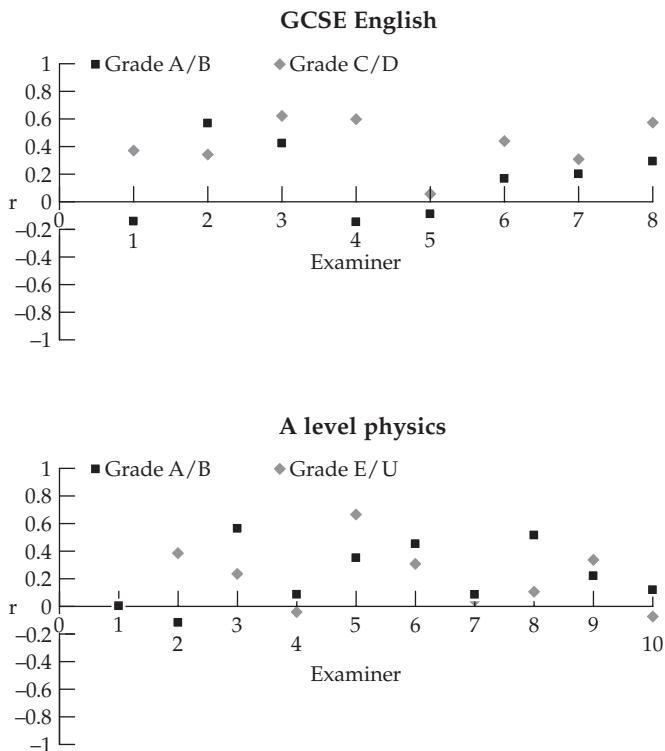
For weak criterion referencing to be acceptable, we must be able to trust that the qualitative judgements made by examiners are reliable and fair. We have already seen that they do not adequately compensate for changes in difficulty of the question paper, but there are other reasons to question them. Associated with the weak criterion referencing approach is the idea that examination standards are embodied in examples of students' work, in relation to the question papers. Thus, following Dearing (1996), enormous resources have gone into creating a national archive of question papers and examples of students' performances, so that the standards over the years and in different subjects can be evidenced. Indeed, part of the standard-setting process has long involved reference to candidates' work on the cut score in the previous year.

Baird (2000) investigated whether these exemplars influenced examiners' judgements in A level psychology and English by manipulating the exemplars provided to the examiners in an experiment conducted outside the operational grading process. She found that it made no difference whether examiners were given the correct exemplar for grade E or were deceived by being supplied with an exemplar for grade D. Some of the examiners were given no exemplars at all and they still set standards comparable with the other groups. Therefore, it has to be concluded that examiners are setting standards with reference not to these exemplars that they are being supplied with, but with reference to their own mental models of the standard. There is also evidence that examiners are unduly influenced by the consistency of

candidates' performances (Scharaschkin & Baird, 2000). This is an illegitimate effect because candidates are allowed to compensate for weak performances in one area with stronger performances in another in the A level and GCSE examinations. Further, examiners demonstrate a tunnel-vision effect in their judgements, as they make more severe judgements of candidates' work when they judge each question paper independently than when they judge all of their work for A level (Baird & Scharaschkin, 2002).

Weak criterion referencing also relies upon examiners being able to make qualitative distinctions between candidates' work on adjacent marks. Baird & Dhillon (2005) conducted studies with GCSE English and A level physics examiners, asking them to rank-order candidates' work in the seven-mark range in which examiners normally scrutinise candidates' work for a grade boundary decision (see Figure 3). Care had been taken to ensure that the marking of the work included in the study was accurate. Correlations between each examiner's rank-ordering and the marks were low to moderate, and none of the 36 correlations calculated were statistically significant (following correction for multiple testing). None of the examiners rank-ordered candidates' work well for both grade boundaries included in the study. Using a different methodology, Forster (2005) found similar results in business studies, English and geography.

Figure 3 Correlations between examiners' rank-orderings of grade-worthiness and mark



This should not be interpreted as meaning that senior examiners do their job badly. On the contrary, they are selected because they are the best people for the job and show a great deal of diligence in marking and grading candidates' work in the interests of fairness. The task of judging to a precise mark, at the boundary between one grade and the next, is impossible. Candidates can reach that mark through thousands of different routes through the question paper (see Scharaschkin & Baird, 2000). Examiners are expected to be able to make a judgement about the extent to which the performances they see on the question paper are caused by a change in the question paper or in candidate preparedness. Taking these features together, there is no prototypical performance that examiners can look out for – the candidates may have reached their mark by a different, but equally valid, route or the question paper may have enhanced or detracted from their performance.

Nonetheless, if we wish to adopt the weak criterion referencing approach to standards, we would have to accept the inconsistency in results produced (statistically speaking, of course). This would entail not using school performance tables and not using the examination outcomes to measure the health of the education system as a whole. As we have seen in the case of the AEB A level English examination, it is unlikely that education stakeholders in England would be content with the variations in statistical outcomes this definition produces, and would resort to statistical definitions of examination standards to challenge it. This leads us to the next definition of examination standards.

2.5 Conferred power definition

Under this definition, society empowers certain individuals to make judgements regarding where the examination standards lie (Cresswell, 1996; Wiliam, 1996b). The process by which these judgements must be made, the information to be taken into account and the criteria for selection of the individuals to make the judgements are all important for this definition. Once the individuals are appointed, as long as due process is followed, there can be no recourse to appeal against their judgements, which are a speech act (Searle, 1969). Once a speech act has been uttered, it makes no sense to question whether it is true (e.g. 'I pronounce you man and wife' uttered by a priest following a legally conducted marriage ceremony). Under this model, there is no pretence that there is an objective way in which standards can be set – we simply accept the umpire's decision. The 'umpire' could be an examiner, a statistician, an examining board chief executive or anyone who fits the selection criteria defined in the due process. No guarantee about *what* the standard is comes from this definition – it is simply a value judgement that does not necessarily ascribe properties to the objects being judged (Cresswell, 1996).

Baird *et al.* (2000) argued that the due process in the case of examination standards is underspecified because the weight that should be given to different sources of information is not given. (Note that the 'conferred power definition' was termed the 'sociological definition' in Baird *et al.*, 2000.) This leaves the pronouncements about standards open to appeal and standard-setting does not operate as a speech act in practice.

Three significant examination crises have occurred in recent years in the UK:

1. the Scottish Qualifications Authority's (SQA's) problems in releasing examination results on time in 2000
2. Edexcel's problems in delivering examination results in 2001
3. concerns regarding examination standards at A level in the English examining boards, particularly at OCR, in 2002.

In an important paper, McCaig (2003) points out that these crises were linked to government policies and resulted partly from inadequate time for examining boards to deliver the government's objectives: integration of vocational and academic examining boards (applies to all three crises), reduction in the number of examining boards (applies to the 2001 and 2002 crises in particular) and delivery of new qualifications (all three crises). McCaig argues that the government distanced itself from the first two crises, but was not so successful with the third. After all, the then Secretary of State for Education and Skills, Estelle Morris, resigned not long after, stating her lack of capability as one of the reasons.

Throughout what became known as the 2002 examinations fiasco, fascinating statements were made by various parties regarding who had ownership of the standards. Allegations were made that the examination standards had been 'fixed' by examining board chief executives, particularly the OCR chief executive, Ron McLone. The chairman of QCA, William Stubbs, and even Estelle Morris herself were accused of influencing the results. Interference from any of these parties was deemed entirely illegitimate by the media. With the exception of cohort referencing, only the conferred power definition of examination standards could be delivered by chief executives of the examining boards, the chairman of QCA or the Secretary of State for Education and Skills. Sir Mike Tomlinson's review cleared QCA and the Secretary of State of any interference and asked the examining boards to look again at the grading of some of the examinations.

Setting aside the 2002 issues per se, the accountability for examination standards is interesting. If the chief executives of the examining boards cannot adjudicate, who is qualified? Delegation of the standards to the 500 committees is a recipe for chaos, with no co-ordination of the standards, policies and approaches being taken. However, the Accountable Officer at each examining board has a formal role within the QCA's code of practice for examinations. Questioning of their role was a rejection of the notion that there is anything further to examination standards than weak criterion referencing. Arguably, the government should have no role in the setting of examination standards, as it should be for educational assessment experts to decide this for society and it should not be a political matter.

Awarding bodies and QCA accepted that QCA has a role in defining the examination standards (House of Commons Education and Skills Committee, 2003). As mentioned previously, following summer 2002, weak criterion referencing was given more emphasis and Chairs of Examiners were elevated to a new role in the QCA's code of

practice – changes to their recommendations could not be made without going through new procedures, which could ultimately result in a public wrangle between the examining board, the Chair of Examiners and the regulator. However, the nature of the accountability of Chairs of Examiners, Accountable Officers, QCA and the DfES and the relations between them have never been specified in a Memorandum of Understanding, as proposed by Tomlinson (2002).

The next section argues that it is not possible to specify the due process completely without making strange decisions in some instances. The conferred power perspective implies a trust in experts that has long since ceased to be a feature of UK society. Accountability is now an integral part of our education system and it would be difficult to envisage a conferred power definition being adopted (Broadfoot, 1996).

3 Construction of an examination comparability preference model

Distinctions can be made between the setting and maintaining of examination standards, as the bar has to be fixed at a particular height in the first year and questions about whether the bar is at the correct height do not necessarily have to be addressed in subsequent years. The previous definitions have been related to the setting of examination standards, although they also apply to the maintenance of examination standards because in practice standard setting and maintenance are similar, with questions about the appropriateness of the bar height being raised fairly regularly and considered to be legitimate. Comparability of examinations can be seen as distinct from standard-setting altogether and comparability issues are often dealt with as part of a research exercise, outside of the standard-setting process. Nonetheless, the definitions of examination standards are integral to examination comparability research and, in practice, comparability issues are part of the considerations during awarding.

Particularly when new examinations are introduced, there is controversy about whether examination standards have been maintained and reference is made to different sources of information supporting different definitions of examination standards. This was evident in the introduction of the Curriculum 2000 examinations, with arguments being made in favour of more reliance upon examiners' judgements. All of the definitions of examination standards outlined are common currency in the debates surrounding the release of examination results. Baird *et al.* (2000) argue that examining boards have to gauge the values that are acceptable to education stakeholders and set standards in that context. After all, these values reflect society's expectations regarding examination standards because of the way in which the results will be used. Standards, they argue, do not exist in any objective sense because the effect of candidate performances and the difficulty of the examination paper cannot be disentangled.

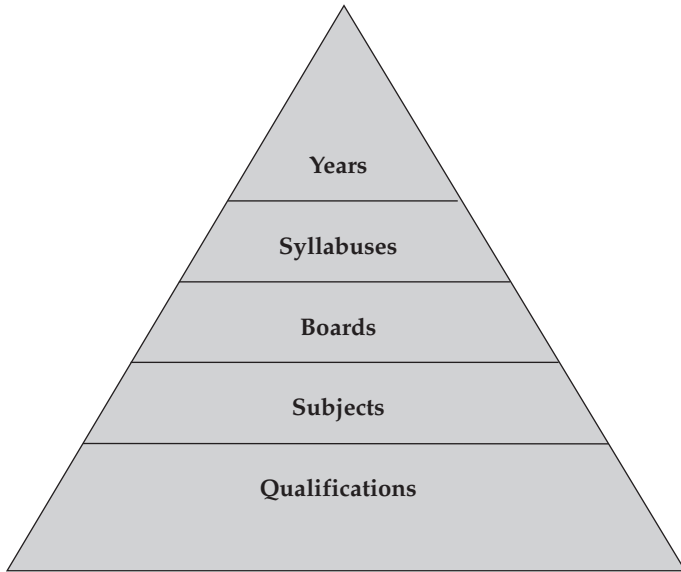
More recently, Newton (2005) argued for a diktat model, in which, for the sake of clarity about educational standards, a single definition is adopted. In particular, Newton argued that a linking construct should be defined whenever standards are set, so that it can be made clear upon what basis the standard is said to be

comparable. He does not select a particular definition to favour, but dismisses the approach presented in Baird *et al.* (2000) as unsatisfactory (terming it the ‘contest’ approach) because it is unclear about what is being maintained. So could we choose to prioritise a particular definition, making clear the construct that is being maintained? Given the discussion surrounding the definitions presented in this chapter, let us consider the implications of the diktat model.

Unfortunately, the situation is not currently even as simple as choosing a particular construct and linking test standards along that construct because examination standards are expected to be maintained in a variety of ways too (Figure 4). Even using a single definition, the evidence regarding comparability for these different things to be compared can be contradictory. For example, the catch-all definition could provide evidence that standards had been maintained between years for a particular syllabus, but that it was no longer comparable with another syllabus in the same subject.

Extending Newton’s diktat model, we could also specify not only the definition of examination standards and the construct to be linked, but our priorities for examination comparability; so that it is clear what information should be given preferential treatment. An alternative under the diktat model would be to select only one of the types of possible comparability, most likely between years. Note that this is not the only possibility under the diktat model, as Newton does not specify which construct should be linked, so it is open for use in different systems. Setting up the diktat model with a link between years would explicitly reject any notion that the examination system had any responsibility for comparability of standards between examining boards, between subjects or between qualifications. Therefore, no points systems in which grades are aggregated across qualifications would be fair – undermining the UCAS and school performance table points systems. Given that these represent two important purposes of UK educational examination assessment, it is easier to envisage that we could retain the expectations about what kinds of comparability should be maintained, but prioritise them. An attempt to make one such priority list is given in Figure 4, with items higher up the pyramid being preferred over those lower down. A logical approach like this is consistent with Donald Schön’s (1988) ‘Technical Rationality’ model of decision making, in which rules and preferences can be written in a logical system. Assessment specialists’ technical knowledge could be used in this technical-rationality approach to formulate rules for deciding what kind of comparability should be preferred over others.

In general, comparability between years is given supremacy in operational standard setting, as those sitting the examinations and teaching the syllabus are entitled to continuity between years in their expectations about the standards. Otherwise, teaching and learning would be problematical, as the results would give little feedback about expectations of students’ performances in the examination. However, the Joint Council for Qualifications has a programme of statistically screening for possible discrepancies between examining boards’ standards in the same subject. Where a subject appears to have been out of line with other examining boards, would it make sense to continue to prioritise between-year comparability for that syllabus?

Figure 4 A possible examination comparability preference model

Equally, if two syllabuses in the same subject appeared to have discrepant standards, would it be correct to continue with that situation? Under these circumstances, the order of priority of years, syllabuses and boards may be changed on a rational basis. But this situation could simply be seen as an exception to the general model and logical guidance could be given, such as:

prioritise between year comparability, except

if

specific information is presented showing that between syllabus/board comparability is discrepant,

then

prioritise between syllabus/board comparability.

This would be fine if there were no circumstances under which this would be irrational, but such circumstances do arise in practice. For example, syllabuses have to be classified into subject areas to make comparisons between them. There may be good reasons to question such classifications, as they implicitly set up comparability expectations that may not be legitimate. For example, are philosophy and critical thinking syllabuses in the same subject area? If there is a question mark over any aspect of the comparison, it would make sense to continue to prioritise comparability between years.

Also, if philosophy was too easy compared with critical thinking using a value-added definition of examination standards, but philosophy was much harder than

other A level subjects using a common-candidates definition of examination standards, would it be sensible to adjust the standard of the philosophy examination to make it tougher? This is a value judgement, the answer to which will differ between individuals and in different contexts – it can be prescribed, but only if we are willing to accept situations such as philosophy being aligned with critical thinking (statistically speaking), but being made even tougher compared with all other A levels.

Annually, standard-setting examples arise in which different features of examination comparability are given priority:

- GCSE French, German and Spanish have very similar assessment structures and there are expectations *regarding inter-subject commonalities in boundary marks* for certain assessments, even though this may disrupt comparability between years to some extent. Prioritising commonalities in boundary marks between other qualifications would have disastrous effects upon between-year comparability and this would not be given priority elsewhere in the system.
- In the merger between AEB and NEAB, in which AQA was formed, for a number of years priority was given to *between-board comparability* rather than between-year comparability. Attaining similar standards across the new examining board was deemed more important than between year comparability, as the pressure from the government to have fewer examining boards was in part to bring about greater consistency in standards. (In practice no large changes in standards were made in any one year.)
- When the Applied GCEs were introduced in summer 2006, statistical information was used in the standard-setting process that would help to align the standards of the Applied GCEs with their respective academic GCEs. However, it quickly became apparent that the value-added between GCSE and Applied GCE was different from the value-added between GCSE and academic GCE (similar patterns emerged, subject by subject across the examining boards). As the Applied qualifications were designed to be different from traditional GCEs, it would be nonsensical to ignore those differences and expect the same value-added from each type of qualification. In this example, with standards being set for the first time, although a great deal of information was considered from different sources, eventually the decisions reached reflected between-qualification comparability, in terms of the weak criterion referencing approach.

These cases show the kinds of contextual, education stakeholder values that the examining boards and their senior examiners take into account to set the examination standards. To do otherwise, would be to fly in the face of society's expectations of the examination standards. A diktat model is only possible to the extent that it is possible to do this. In practice, under some circumstances, particular standards definitions and types of comparability will come to the fore, but these may change depending upon the context.

Schön (1988) argues that there are features of the real world that make the Technical Rationality model of decision making difficult:

- it is complex, making it difficult to ascertain which features to focus upon
- it is unstable, making it difficult to generate heuristics
- unique instances arise for which the professional has no real reference point
- conflicting values mean that not all constraints can be satisfied (see Table 4).

Major theoretical advantages of the diktat model are its transparency about what standard is being maintained and its scientific, logical approach, but its major drawback is that it would not provide society with the qualifications it currently desires: certain expectations of comparability would be deemed beyond the realms of possibility. Of course, the contest model does not guarantee that these forms of comparability are delivered, but neither does it reject many of them as irrelevant. Even if the diktat model was adopted, it would soon falter, as no attempt would be made to address the concerns of various educational stakeholders. Naturally, education stakeholders adopt different positions regarding examination standards and comparability depending upon their perspectives, and this is not singular even for individuals over time. Examination standards are highly politicised because they are intimately related to beliefs about individuals' and societies' economic prospects (Wolf, 2002; McCaig, 2003).

Nonetheless, some argue that separate credit systems should be set up to deal with linking of examination outcomes in various post hoc ways, such as scaling the examination grades to compensate for differences in difficulty for a university entrance points system. Grading itself should only deal with the originally envisaged linking construct. Australia's university entrance system is based upon students' examination grades, but scaling for differences in difficulty between states' examinations is carried out. Lamprianou discusses this further in his commentary on Chapter 9.

There is no doubt that this is a theoretically attractive way out of the quagmire of comparability as defined in the English examination system, but would the social and political educational structures support such a change? Persuading stakeholders to give up closely valued features of the English grading system would be difficult, even if they are not believed to be well-delivered currently, due to the tensions in definitions of examination standards in use outlined earlier, and because the technicalities of doing something different would be seen as obfuscatory. But this is not an argument for the status quo, it is a statement of the realities that will face any such move. This is not just a technical matter – it is culturally and politically embedded.

Table 4 Features of the real world that make professional, scientific decision making impossible

Feature of real world	Example in the standard-setting task
Complexity	Different forms of comparability. Contradictory evidence regarding different kinds of comparability. Differences between assessment formats. Multiple performance attributes in candidates' work. Changes in the content of assessment.
Instability	Political and educational stakeholder values change. Difficulty of the examinations changes. Statistical relationships between predictor variables and examination outcomes change.
Uniqueness	Candidates respond to assessments in novel ways.
Value conflict	Different definitions of standards. Different perspectives of educational stakeholders.

4 Conclusion

What counts as a definition of examination standards – what do we expect from such a thing? Newton (2005) argues that the important feature of a definition of examination comparability is the construct being used to link the standards. Whilst important, this in itself does not provide the whole definition. Test constructors have in mind a particular construct when they create the tests, but they do not have control over the myriad ways in which society expects comparability to be maintained. These comparability expectations are not inherent to the tests themselves either – they are externally, and often subsequently, imposed requirements. School performance tables were introduced with a credit system expounding comparability between all GCSEs, GNVQs and other qualifications at the same level.

Empirically, it is easy to show the faults in these credit systems, but if they work well enough as a currency system, then pragmatics entail that whatever can be done to root out the worst cases of lack of comparability must be done. So a test may be designed to assess ability in science and two tests can be linked using this construct, but these tests may also be linked with other tests, using different constructs, such as 'general academic ability'. In any case, any particular examination does not assess a single construct and can be linked on more than one theoretical (after all, they are all theoretical) construct. Indeed, different assessments for any particular qualification typically have low correlations, implying that different constructs are being measured.

A definition of examination standards should ideally meet the following criteria (adapted from Fawcett's (2005) criteria for evaluation of theories):

1. It should have a theoretical underpinning, referring explicitly to the educational intentions of standards and comparability. The theory should be consistent, as opposed to predicting more than one outcome for any particular case.

2. The definition should be testable and supported by evidence.
3. As with any good theory, the definition should be parsimonious.
4. The definition should be practically useful in our educational culture.
Contributions to academic debate are useful theoretically, but ideally a definition would conform to this criterion.

As the foregoing discussion shows, the author's view is that all of the definitions of examination standards have weaknesses and this is also true in relation to the criteria above. Chapter 11 looks at what developments are needed in this area to strengthen our definitions of examination standards. All of the definitions are open to empirical test, most of them are weak on theoretical underpinnings and there are serious problems with the practicality of some of them in the English educational, cultural context. Moving away from a balance between the weak criterion referencing definition and the statistical, catch-all definition – i.e. away from the contest model – would require a radical shift in England's educational culture.

Expectations about examination standards and comparability exist because of the way examination results are used in society. Prioritising a particular definition would make the examination results less useful, although Newton (2005) argues that assessment specialists have gone too far in trying to achieve the impossible. A single definition would not solve many problems either. Not only are there tensions between different definitions, there are tensions within them, with competing approaches being used to try to measure examination comparability throwing up different results and linking between different kinds of qualifications suggesting different conclusions. Hence, the lively debate in the following chapters and with the respondents.

References

- Aldrich, R. (2000). Educational standards in historical perspective. In H. Goldstein & A. Heath (Eds.), *Educational standards* (pp. 39-56). Oxford: Oxford University Press for The British Academy.
- Baird, J. (2000). Are examination standards all in the head? Experiments with examiners' judgments of standards in A level examinations. *Research in Education*, 64, 91-100.
- Baird, J., Cresswell, M.J., & Newton, P. (2000). Would the *real* gold standard please step forward? *Research Papers in Education*, 15, 213-229.
- Baird, J., & Dhillon, D. (2005). *Qualitative expert judgements on examination standards: Valid, but inexact*. Internal report RPA 05 JB RP 077. Guildford: Assessment and Qualifications Alliance.

Baird, J., Ebner, K., & Pinot de Moira, A. (2003, October). *Student choice of study in Curriculum 2000*. Paper presented at the International Association for Educational Assessment Annual Conference, Manchester.

Baird, J., & Jones, B.E. (1998). *Statistical analyses of examination standards: Better measures of the unquantifiable?* Research Report RAC/780. Assessment and Qualifications Alliance.

Baird, J., & Scharaschkin, A. (2002). Is the whole worth more than the sum of the parts? Studies of examiners' grading of individual papers and candidates' whole A-level examination performances. *Educational Studies*, 28, 143–162.

Bell, R., Cardello, A.V., & Schutz, H.G. (2005). Relationship between perceived clothing comfort and exam performance. *Family and Consumer Sciences Research Journal*, 33, 308–320.

Broadfoot, P.M. (1996). *Education, assessment and society: A sociological analysis*. Buckingham: Open University Press.

Cresswell, M.J. (1996). Defining, setting and maintaining standards in curriculum-embedded examinations: Judgemental and statistical approaches. In H. Goldstein & T. Lewis (Eds.), *Assessment: Problems, developments and statistical issues* (pp. 57-84). Chichester: John Wiley and Sons.

Cresswell, M.J. (1997). *Examining judgments: Theory and practice of awarding public examination grades*. Unpublished PhD thesis, University of London Institute of Education.

Cresswell, M.J. (2000). The role of public examinations in defining and monitoring standards. In H. Goldstein & A. Heath (Eds.), *Educational standards* (pp. 69-104). Oxford: Oxford University Press for The British Academy.

Day, J.A. (1992). *652 English literature 1991. Reconvened grade awarding meeting report*. Unpublished internal paper, Associated Examining Board.

Dearing, R. (1996). *Review of qualifications for 16–19 year olds*. London: School Curriculum and Assessment Authority.

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

Department for Education and Skills. (2002). *The national literacy and numeracy strategies. Including all children in the literacy hour and daily mathematics lesson. Management guide*. DFES 0465/2002. London: Department for Education and Skills.

Department for Education and Skills. (2006). *Specialist schools information on the standards site*. Available at <http://www.standards.dfes.gov.uk/specialistschools/>

Dowie, J., & Elstein, A. (1988). *Professional judgment. A reader in clinical decision making*. Cambridge: Cambridge University Press.

Erwin, P.G. (1999). Attractiveness of first names and academic achievement. *Journal of Psychology: Interdisciplinary and Applied*, 133, 617–620.

Fawcett, J. (2005). Criteria for evaluation of theory. *Nursing Science Quarterly*, 18, 131–135.

Forrest, G.M., & Shoesmith, D.J. (1985). *A second review of GCE comparability studies*. Manchester: Joint Matriculation Board on behalf of the GCE Examining Boards.

Forster, M. (2005). *Can examiners successfully distinguish between scripts that vary by only a small range of marks?* Unpublished internal paper, Oxford Cambridge and RSA

Gillbourn, D., & Youdell, D. (2000). *Rationing education. Policy, practice, reform and equity*. Buckingham: Open University Press.

Gilmore, A. (2002). Large-scale assessment and teachers' capacity: Learning opportunities for teachers in the National Education Monitoring Project in New Zealand. *Assessment in Education*, 9, 343–365.

Goldstein, H. (1986). Models for equating test scores and for studying the comparability of public examinations. In D.T. Nuttall (Ed.), *Assessing educational achievement* (pp. 168–184). London: Falmer.

Good, F.J., & Cresswell, M.J. (1988). Grade awarding judgments in differentiated examinations. *British Educational Research Journal*, 14, 263–281.

House of Commons, Education and Skills Committee. (2003). *A level standards*. Third report of session 2002–03. HC 153. London: The Stationery Office Limited.

Impara, J.C., & Plake, B.S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions of the Angoff standard setting method. *Journal of Educational Measurement*, 35, 69–81.

Independent Appeals Authority for School Examinations. (1991). *Independent Appeals Authority for School Examinations annual report*. London: Independent Appeals Authority for School Examinations.

Joseph, K. (1984). Speech by Sir Keith Joseph, Secretary of State for Education and Science, at the North of England Conference, Sheffield, Friday 6 January, 1984. *SEC annual report 1983–1984*. London: Secondary Examinations Council.

- Kingdon, M.J. (2005). *Tomlinson revisited. A supplement to occasional paper No. 1*. Hellingly, East Sussex: The Examination on Demand Assessment Advisory Group.
- Lane, A.M., Whyte, G.P., Terry, P.C., & Nevill, A.M. (2005). Mood, self-set goals and examination performance: The moderating effect of depressed mood. *Personality and Individual Differences, 30*(1), 143–153.
- Lightfoot, L. (2006, May 10). Modular exams 'damaging degree courses'. *Daily Telegraph*.
- Mansell, W. (2006, June 9). Pick A, B or C for a GCSE. *The Times Educational Supplement*, p. 1.
- Martin, D. (2005). *Report on the 2004 scholarship to the Deputy State Services Commissioner by the review team led by Doug Martin*. New Zealand: State Services Commission.
- Massey, C., & Baird, J. (2001). *A comparison between practical coursework routes in GCE chemistry (0654)*. Internal Report RC/140. Guildford: Assessment and Qualifications Alliance.
- McCaig, C. (2003). School exams: Leavers in panic. *Parliamentary Affairs: Special Issue, 56*, 471–489.
- Newton, P.E. (2005). Examination standards and the limits of linking. *Assessment in Education, 12*, 105–123.
- Pinot de Moira, A. (2002). *Preliminary analysis of the summer 2002 A level results*. Internal paper, RC/188. Guildford: Assessment and Qualifications Alliance.
- Raudenbush, S. (1994). Searching for a balance between a priori and post hoc model specification: Is a 'general approach' desirable? *School Effectiveness and School Improvement, 5*(2), 196–198.
- Scharaschkin, A., & Baird, J. (2000). The effects of consistency of performance on A level examiners' judgements of standards. *British Educational Research Journal, 26*, 343–357.
- Schön, D.A. (1988). From technical rationality to reflection-in-action. In J. Dowie & A. Elstein (Eds.), *Professional judgment: A reader in clinical decision making*. Cambridge: Cambridge University Press.
- Searle, J.R. (1969). *Speech acts. An essay in the philosophy of language*. Cambridge: Cambridge University Press.
- Taylor, M. (2006, August 1). GCSE coursework to be curtailed to stop internet cheats. *The Guardian*.

Tomlinson, M. (2002). *Inquiry into A level standards. Final Report*. London: Department for Education and Skills.

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

Wiliam, D. (1996b). Standards in examinations: A matter of trust? *The Curriculum Journal*, 7, 293–307.

Wilmot, J., & Rose, J. (1989). *The modular TVEI scheme in Somerset: Its concept, delivery and administration*. London: Report to the Training Agency of the Department of Employment.

Wolf, A. (2002). *Does education matter? Myths about education and economic growth*. London: Penguin.

Working Group on 14–19 Reform. (2004). *14–19 curriculum and qualifications reform (Tomlinson Report)*. London: Department for Education and Skills.

Zahr, L. (1985). Physical attractiveness and Lebanese children's school performance. *Psychological Reports*, 56, 191–192.

COMMENTARY ON CHAPTER 4

Harvey Goldstein

Jo-Anne Baird quotes me in 1986 as advocating a cohort referenced system. This is correct, and while I would not necessarily advocate this now in its pure form, I do not think it can be so summarily dismissed as in this chapter.

My argument was that the system would be transparent and as such throw the responsibility for interpretation upon the users of examination results, and in particular it would eschew the notion that present procedures can provide objective and fair comparisons. Baird argues that such a system could not be used by government and others to measure trends over time. In fact the present system cannot do that satisfactorily either, despite claims to the contrary (see for example Goldstein, 2000). A system that avoids the ill-informed debates that take place every year when examination results are published is surely to be welcomed. As she points out, specially developed measures would be needed, for example to evaluate policy changes, and the development of these would be welcome.

In fact, something akin to a cohort referencing system applies to university degree classes, where each university decides on its allocation with a minimal attempt to ensure any kind of comparability, even over time within institution. It works, partly, because of the reputation built up over time attached to each institution, of which users are well aware. It is simply not justified to claim that a similar system applied to examinations would lead to dumbing down. Nor do I see why new qualifications would necessarily 'struggle for recognition'. There are all kinds of ways of bringing innovations into the curriculum and examination system.

I can see, however, that for small-entry subjects, initial difficulties could be serious. Thus, if the suggestion were ever acted upon I would suggest that it first is applied on an experimental basis to mass entry subjects such as English and mathematics. Where I do agree strongly with Baird is in her view that such a system would stand little chance of being adopted in the current climate.

Reference

Goldstein, H. (2000). Discussion (of the measurement of standards, by David Bartholomew). In H. Goldstein & A. Heath (Eds.), *Educational standards* (pp. 138-149). Oxford: Oxford University Press for The British Academy.

COMMENTARY ON CHAPTER 4

Robert Coe

In this commentary on Jo-Anne Baird's chapter, I attempt to do three things. Firstly, to try to clarify some fundamental conceptual distinctions in the different meanings of 'comparability'. Secondly, to argue that one particular conception of comparability, construct comparability, is under-recognised in Baird's chapter (and elsewhere), but it provides an important perspective for understanding comparisons of different examinations. Thirdly, to explore the relationships between different views about the interpretation of examination grades or the uses to which they may be put, and the different conceptions we may have of comparability.

Conceptualising 'comparability'

Baird's chapter on comparability points out some of the anomalies that arise from popular understandings of the notion of 'standards' and goes on to describe and evaluate five specific 'definitions of examination standards from assessment specialists'. In doing this she illustrates convincingly the practical problems of each definition. However, she presents no overall conceptual framework for thinking about different meanings of 'comparability', and the reader may be left with some fundamental questions. Are these different definitions merely different operationalisations of the same fundamental conception of what 'comparability' means, or are they conceptually different? Does each definition represent a pure conception, or are some effectively hybridisations arising from the mixing of ideas? In each case, what does 'comparability' actually mean?

Much existing thinking about comparability issues within the UK has focused on the processes by which test scores are translated into an interpretable 'standard' (e.g. Wiliam, 1996a). There seems to be a broad consensus that there are basically two ways one can do this: the standard is either specified in terms of performance criteria, or in terms of statistical norms for some population. The terms *criterion-referenced* and *norm-referenced* are widely used to describe these two approaches, though as Wiliam (1996a) makes clear, the distinction is somewhat problematic in practice: 'a criterion-referenced test is just a well-designed norm-referenced test that has had the luxury of being restricted to a very small domain' (p. 295). A similar idea seems to underlie Jaeger's (1989) use of the terms *test-centred* and *examinee-centred* to distinguish between approaches to setting a standard that consider only features of the test and those that take into account the performance of examinees. The terms *performance comparability* and *statistical comparability* seem to capture this distinction.

In her chapter, Baird appears to adopt this broad dichotomy, describing two of the approaches, *cohort-referencing* and the *catch-all definition*, as coming from 'the statistical camp of assessment literature definitions of examination standards', and

two others, *criterion-referencing* and *weak criterion-referencing*, where the standard resides in the observed (and evaluated) test performance, regardless of how many candidates achieved it. It is important to remember that the distinction between these two kinds of comparability can only be maintained at the level of idealisations. A pure *performance comparability* view would require us to judge the standard of a candidate's test performance by considering only the test, the context in which it was taken and the candidate's responses to it, but without any knowledge of how any other candidate had performed on that test – or even on any similar tests. Baird provides comprehensive, well-illustrated and convincing arguments, however, that in practice this cannot be satisfactorily done. On the other hand, a pure *statistical comparability* approach would compare the standards of different examinations using statistical information about how candidates with particular characteristics have performed on them, but without any knowledge of what those candidates were actually required to do. Again, Baird shows that such a method is unlikely to be satisfactory in practice.

Nevertheless, these idealisations are important. If 'comparability' can be understood in theoretically different ways, then terms such as 'standards' or 'difficulty' may also have more than one meaning. Claims such as those by Chris Woodhead that examinations are getting easier (Box 5 in Baird's chapter) may not actually contradict the claims of other studies (McGaw *et al.*, 2004) that they are not. The bases for, and meanings of, these claims are quite different. Characterising this debate as a 'tug of war' between progressives wanting to modernise the curriculum and traditionalists wanting to preserve elitist selection may be reading too much into their differences; they may simply be using the same word to mean two quite different things.

Baird presents a further approach to standard-setting, the *conferred power* definition, which does not fit into either the *performance* or the *statistical* conceptualisation of comparability. This approach sees standards as a pure social convention, defined by the values of a 'community of practice' rather than by any explicit rationale (Wiliam, 1996b). If our goal is to try to understand what is meant by 'comparability', however, then the *conferred power* definition can be dismissed fairly readily, since it offers nothing in the way of a conceptualisation. Of course, it is true that expert judgement and the application of subjective values are required to set standards. It is also true that some degree of trust in the judgements of 'awarders accepted as competent to make such judgements by all interested parties' (Cresswell, 1996, p. 79) must be a requirement of any system. However, this definition tells us nothing about how such trust might be established – or rebuilt if it is lost – or how these awarders come to be 'accepted as competent'. The *conferred power* definition offers no better answer to the question of why one examination is, or is not, comparable to another than 'Because I say so'. Such an answer seems unlikely to convince critics such as Woodhead or Dunford (cited by Baird) that they must simply take it on trust.

The case for 'construct comparability'

How, then, do we operationalise comparability? If the *conferred power* definition of comparability is not really a definition at all, and the *performance* and *statistical*

definitions provide useful conceptual idealisations, but have limited practical value, what are we left with? Fortunately, there is an alternative conceptualisation of comparability, *construct comparability*, which is both logically coherent and practically operationalisable.

The concept of *construct comparability* arises from a perspective of trying to understand what it means for two examinations to be compared, rather than trying to define the meaning of a 'standard'. Logically, for a comparison between two things to be meaningful, there must be something they have in common, in terms of which they can be compared. A comparison has no meaning unless it relates to the amount or quality of some construct. In the context of comparing examination standards, it follows that if we can identify some common construct, shared by two or more examinations, then we have a basis for judging whether they are 'comparable'. This idea is developed further in my own chapter on common examinee methods (Chapter 9), where a number of examples of analyses are described whose results can be interpreted in terms of *construct comparability*.

It is possible to see *construct comparability* as subsuming both *performance* and *statistical* conceptions. The whole idea of a criterion-referenced standard arguably depends on identifying a particular level of some construct that can be defined sufficiently precisely. Without some such construct in mind, we cannot say that one criterion would be harder to meet than another. Hence *criterion-referencing* may be seen as a special case of *construct comparability*. It is also arguable that at least some forms of what appear to be *statistical comparability* are actually *construct comparability*. Although it is always possible to make statistical comparisons of the grades achieved in different examinations by 'comparable' candidates, it makes little sense to do so unless some theoretical construct guides the choice of the basis on which candidates are seen as 'comparable'. The mechanism by which a set of starting characteristics can be converted into examination grades in similar ways across different examinations seems to require some common construct to link them if the comparison is to be meaningful. From this, it seems tempting to conclude that, just as 'construct validity is the whole of validity' (Loevinger, 1957, p. 636), perhaps *construct comparability* is the whole of comparability.

Of course, this idea is not new. Wiliam (1996b) actually uses the term 'construct-referenced' assessment to account for the fact that a group of assessors may agree about the standard of a piece of work, even where there are no explicit criteria against which to judge it. They may nevertheless share an understanding of a broad construct which he calls 'levelness' but which might be interpreted as 'English attainment'. An even more explicit presentation of the idea of *construct comparability* can be found in Newton (2005) who discusses how a 'linking construct' can be used to establish the comparability of a group of examinations. This idea is discussed further in Chapter 9 of this volume.

Comparability in relation to interpretation and use of examination grades

From a *construct comparability* perspective, we can compare two or more examinations

only if a common construct has been identified. However, just as the same examination may be interpreted in different ways for different purposes, there may be some cases in which more than one possible construct could be used as a basis for comparing the same set of examinations. It follows that there may be more than one view about their comparability: in terms of construct 'A', examination 'X' may be judged 'harder' than 'Y', but in terms of construct 'B' the position would be reversed.

The fact that there are multiple uses for examinations and multiple possible interpretations of their results implies that there may be multiple possible constructs that could be used to define comparability. Realistically, therefore, it is unhelpful to talk about comparability of examinations unless we are clear about the particular purpose for which we want to use and interpret those examinations.

The issue Baird raises about prioritising comparability across years versus comparability across syllabuses, etc., is a secondary one. If we could agree a construct against which to compare, and were in a position to create examinations from scratch, then we could theoretically achieve comparability for all these comparison groups together. In practice, of course, if we found that existing examinations were not comparable in their 'standards', then there would be a tension between achieving comparability within a particular year and across years. This would be a political rather than a technical problem, however.

To the more fundamental theoretical problem of multiple bases for comparability, there are perhaps three possible responses. The first would be to choose one preferred basis for comparability. This is Newton's (2005) *diktat* model, and would amount to privileging one use/interpretation of examination grades, with the corollary that other uses may then not be valid. The second would be to acknowledge that there are a limited number of valid bases for understanding comparability and adopt some kind of optimisation strategy – or 'contest' (Newton, 2005) – among them. One such has been described by Wiliam (1996b) as keeping a number of needles on a dial out of the red zone, so that no valid judgement of comparability would place different examinations too far from being in line. The price to be paid for this approach is that the meaning of 'comparability' becomes blurred in a pragmatic compromise – politically acceptable, but not rationally defensible. Newton (2005) argues that this is too high a price, and hence prefers the *diktat* model. However, it could be argued that the *diktat* model is just a special case of the *contest* model in which one particular interpretation has won the contest.

There may be a third possibility, however. Whatever process is used in the grade-setting process, it should be acknowledged that there is no absolute, universal sense in which different examinations are comparable; comparability is always relative to a particular use or interpretation. Nevertheless, if examinations are to be used for a particular purpose then we can readily convert, or rescale, their results to make them comparable for this purpose. This may therefore be thought of as a *variable conversion* model. Just as there is not a single conversion rate between currencies at any given time (it depends which market you go to), there is no single conversion rate among examination grades. The conversion rate is variable and depends on the particular

interpretation of those grades and the linking construct that underlies it. Although the complexity and changeability of meaning of 'comparability' implied in such an approach might make it seem politically unacceptable, the fact that Average Marks Scaling has been used in this way in Australia for many years (see Chapter 9 by Coe) suggests that the political problems may not be insuperable. If that is so, it may be that this approach offers a solution to the problem of comparability that is both socially acceptable and conceptually defensible.

References

- Cresswell, M.J. (1996). Defining, setting and maintaining standards in curriculum-embedded examinations: Judgemental and statistical approaches. In H. Goldstein & T. Lewis (Eds.), *Assessment: Problems, developments and statistical issues* (pp. 57-84). Chichester: John Wiley and Sons.
- Jaeger, R.M. (1989). Certification of student competence. In R.L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: American Council on Education/Macmillan.
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3, 635-694.
- McGaw, B., Gipps, C., & Godber, R. (2004). *Examination standards: Report of the independent committee to QCA*. London: Qualifications and Curriculum Authority.
- Newton, P.E. (2005). Examination standards and the limits of linking. *Assessment in Education*, 12, 105-123.
- William, D. (1996a). Meanings and consequences in standard setting. *Assessment in Education*, 3, 287-307.
- William, D. (1996b). Standards in examinations: A matter of trust? *The Curriculum Journal*, 7, 293-307.

RESPONSE TO COMMENTARIES ON CHAPTER 4

Jo-Anne Baird

Response to Harvey Goldstein

Goldstein points out that university degree results are interpreted by selectors, employers and other users of the qualifications, who attach value to those results at least partly on the basis of the reputation of the university awarding them. But degree results vary between subjects and institutions, after controlling for prior attainment (Chapman, 1996), raising questions regarding comparability of standards (Chapman, 1997). With such variability, and users of the qualification results having experience of only a few cases from any individual university department, judgements regarding the value of the qualifications are bound to be unreliable and subject to bias. Selection to university should be as free from bias as possible. For these reasons, it is questionable whether higher education is a model for secondary-level standard-setting to follow.

Although I argue that examination standards do not exist except as social constructs (Baird *et al.*, 2000) in this chapter, I do not go as far as Goldstein in support of exploring cohort referencing. Attempts to address discrepancies in comparability are important – I argue that a judicious balance between experts' views of candidates' performances in particular subjects and statistical analyses of the outcomes are necessary parts of the current examination system, unless we are willing to abandon some of our expectations regarding what the system delivers or deliver them by other means.

Response to Robert Coe

Coe argues that one way of meeting certain expectations would be to adopt something akin to the Australian Average Marks Scaling. In some respects, this is an attractive option because it allows the examination results to perform certain functions and converts them into a different currency for use in other functions. Underlying this approach is an assumption that a single construct can be used to link all of the examination results and Coe hints that this may be the case. Ability is the term normally used for such a construct and the consequence of such a system is that subjects or examinations that depart from that construct may not be treated fairly under that system. Who is to say whether this assumption is better than those made in the Universities and Colleges Admissions Service (UCAS) tariff? The UCAS tariff gives points for each grade for a range of qualifications. Some universities use these points for admissions. Assumptions underlying the system include the value of different qualifications and the relationships between the grades. Certainly, there are assumptions underlying both approaches. In the case of the UCAS system, the assumptions are based upon stakeholders' value judgements regarding the

worthiness of the qualifications, whereas in the Average Marks Scaling system, they are based upon statistical mechanisms that assume that ability underlies the examination results.

Coe's comments on 'construct comparability' are heavily related to Newton's (2005) argument and his notion of variable conversion follows from my discussion of Newton's argument in the conclusion section of Chapter 4. We are in agreement to the extent that different constructs can be conjured to equate different pairs of tests.

Coe is concerned that no conceptual framework is presented in the chapter and attempts to provide one, with reference to 'norm-referencing' and 'criterion-referencing' distinctions and he also refers to other authors' definitions. Coe has misinterpreted the literature in failing to recognise that the weak criterion referencing (Baird *et al.*, 2000) and conferred power definitions (Cresswell, 1996) were new approaches, adding to the previous and conceptually distinct. Each of the approaches outlined in the chapter has a different stance with regard to what needs to be taken into account and what adjusted for when drawing conclusions about comparability of examination standards (Table 1).

Table 1 Different definitions of examination standards

<i>Statistical approaches</i>	<i>Takes account of ...</i>	<i>Adjusts for ...</i>
Cohort referencing	Students' rank order	Nothing
Catch all	Students' grades	Student, teacher and institutional characteristics
<i>Judgemental approaches</i>		
Criterion referencing	Candidates' performances	Nothing
Weak criterion referencing	Candidates' performances and the assessment itself	Difficulty of assessment
Conferred power	Specified by due process	Specified by due process

Comparability methodologies described in this book may be operationalisations of a specific definition (see Table 1), or they may be applicable to more than one definition. Pollitt *et al.*'s chapter on examination demands is clearly linked with the weak criterion-referencing approach, but his theoretical analysis goes further, touching upon curriculum issues. Adams' chapter on cross-moderation discusses a technique that could be used in conjunction with any of the judgemental methods and the same is true of the Thurstone-pairs technique, described by Bramley. As Schagen and Hutchinson point out in their chapter on multilevel modelling, statistical techniques in practice have been impoverished attempts to implement the catch-all definition. Included in this are value-added approaches, common centres' analyses, use of reference tests (discussed in Murphy's chapter) and subject-pairs (discussed in Coe's chapter).

Cizek and Bunch (2007) write,

... we think it is obvious that any standard-setting procedure necessarily requires participants to bring to bear information about both test content and test takers. It would not be possible for a standard-setting participant to make a judgment about the difficulty of an item or task without relying on his or her knowledge or expectations of the abilities of examinees in the target population. Conversely, it would not be possible for a participant to express judgments about examinees without explicit consideration of the items or tasks presented to the examinees.

Cizek & Bunch (2007, p. 10)

For this reason, operationalisations of standard-setting typically involve use of statistical and judgemental approaches. Attempts to classify them conceptually can quickly become confusing when they are compared with what happens in practice because there are few 'pure' approaches. Techniques for comparing examination standards can often be interpreted according to more than one definition too, as outlined above. Therefore, we cannot simply look at methods or artefacts to tell us which definition is in use – we need practitioners and researchers to be more explicit to be sure what definition they had in mind. For that to happen, assessment organisations would need to make explicit their policy positions in advance.

References

- Baird, J., Cresswell, M.J., & Newton, P. (2000). Would the *real* gold standard please step forward? *Research Papers in Education*, 15, 213–229.
- Chapman, K. (1996). Entry qualifications, degree results and value added in UK universities. *Oxford Review of Education*, 22, 251–264.
- Chapman, K. (1997). Degrees of difference: Variability of degree results in UK universities. *Higher Education*, 33, 137-153.
- Cizek, G.J., & Bunch, M.B. (2007). *Standard setting. A guide to establishing and evaluating performance standards on tests*. California: Sage Publications Inc.
- Cresswell, M.J. (1996). Defining, setting and maintaining standards in curriculum-embedded examinations: Judgemental and statistical approaches. In H. Goldstein & T. Lewis (Eds.), *Assessment: Problems, developments and statistical issues* (pp. 57-84). Chichester: John Wiley and Sons.
- Newton, P.E. (2005). Examination standards and the limits of linking. *Assessment in Education*, 12, 105–123.