A level pass rates and the enduring myth of norm-referencing

Paul Newton Director, Cambridge Assessment Network, Assessment Research & Development

Defining standards and comparability

Comparability can be defined as the application of the same standard across different examinations (Newton, 2007, p.9). If so, then, to understand comparability, you need to understand what it means to apply a standard. The meaning of comparability, in the context of A level examining, has been confused for decades because of a lack of clarity over the definition of the A level standard.

The mythology of A level examining goes as follows: standards were norm-referenced, from the 1960s until the middle of the 1980s, after which they became criterion-referenced. This article will argue that A levels have never been norm-referenced, have never been criterion-referenced and have always been attainment-referenced. However, to make this case, these terms need to be defined with some precision. Crucially, quite distinct versions of these definitions can be identified within different contexts, so we need to focus specifically upon the way in which the terms have traditionally been understood in the context of UK examinations.

The idea of an examination standard being [x]-referenced means that it is linked to, or defined in terms of, the [x]. In criterion-referencing, the standard is defined in terms of written criteria, the satisfaction of which results in the award of a particular grade. As understood in the context of UK examinations, this has typically been related to the notion of 'mastery' testing, such that, for the award of a particular grade, students must have demonstrated a certain level of proficiency across each of the sub-domains that comprise a curriculum area, that is, they must have mastered all of the critical elements of the domain. You can imagine this in terms of a student needing to have achieved at least the passing grade on each of the principal sub-domains of mathematics (e.g. number & algebra; geometry & measures; statistics & probability) for the award of a pass overall. The pass would thereby certify that the student had 'mastered' all elements of the mathematics curriculum.[i] Criterion-referencing involves identifying exactly what students can and cannot do, in each sub-domain of the subject being examined, and then awarding: grade A to those who have satisfied all of the grade A criteria across all of the sub-domains; grade B to those who have satisfied all of the grade B criteria across all of the sub-domains; and so on.

Criterion-referencing contrasts with attainment-referencing, in which the standard is defined more holistically in terms of a certain level of attainment. Instead of judging students on the basis of their profile of attainment across sub-domains, in terms of clearly specified performance criteria, they are judged on the basis of their overall level of attainment in the curriculum area being examined. In effect, instead of there being a set of criteria for the award of the overall grade, there is just a single criterion. In practice, the idea of unambiguously articulating this single criterion, at such a high level of abstraction, turns out to be so implausible as to force the examiner to drop any pretence of referencing standards to written criteria. All that can be done is to provide a general indication of the kinds of knowledge, skill and understanding that might well be associated with the award of a particular grade. In UK examinations, attainment-referenced standards are currently exemplified (not defined) through 'performance descriptions' (not 'performance criteria') relating to hypothetical 'typical' students. Attainment-referencing involves ranking students, in terms of their overall level of attainment, and then awarding: grade A to students with a certain level of attainment (i.e. the level at which students were awarded the same grade in previous years); grade B to students with a lower level of attainment; and so on.[ii]

Finally, in norm-referencing, the standard is defined in terms of a particular norm-group. When used in the context of UK examinations, the norm-group is simply the cohort that took a particular examination at a particular time. So the norm-referenced standard simply represents the level of attainment of a particular student in relation to the level of attainment of all other students who sat the examination in question. Importantly, both attainment-referencing and norm-referencing rank students in exactly the same way, on the basis of their overall level of attainment in the curriculum area. All that differs is how standards are set for the award of each grade. Norm-referencing involves ranking students, in terms of their overall level of attainment, and then awarding: grade A to the top X%; grade B to the next Y%; and so on.

The distinction between norm-referencing and criterion-referencing came from the North American literature on educational measurement. Glaser (1963/1994) explained that: "When such norm-referenced measures are used, a particular student's achievement is evaluated in terms of a comparison between his performance and the performance of other members of the group" (p.7). This broad definition resonates somewhat with the UK usage, although it is not identical, since the latter is specific in referring to the award of grades to fixed percentages of each examination cohort, a practice known in the USA as 'grading on the curve'. Nowadays, in the USA and elsewhere, norm-referencing tends to have a more specific definition, which departs even further from the UK usage: "A norm-referenced test (NRT) is a type of test, assessment, or evaluation which yields an estimate of the position of the tested individual in a predefined population, with respect to the trait being measured" (Wikipedia, 2011). For example, results from a particular administration of an IQ test would not indicate how well you performed in relation to others tested at the same time, but in relation to the spread of scores that might be expected within the entire population. Wiliam (1996) proposed that the term 'cohort-referencing' characterises UK usage more precisely; although we will remain with the more conventional term for the remainder of the present article.

By way of summary, each of these definitions has different implications for comparability: norm-referencing specifies that students with the same rank (from their respective examinations) should be

awarded the same grade; criterion-referencing specifies that students with the same profile of proficiency (from their respective examinations) should be awarded the same grade; attainment-referencing specifies that students with the same overall level of attainment (from their respective examinations) should be awarded the same grade.[iii]

The myth

The primary aim of the present article is to dispel a widely-believed myth, which goes something like this:

> *For the first 25 years or so, the maintenance of standards of A-levels relied substantially on the constraints of so-called norm referencing, i.e. a constant proportion of candidates in each subject was each year awarded the same grade. […] it differentiates only between those who took the test at the same time and the results have no validity from one year to another.* (Stubbs, 2002, p.4)

This is a quotation from an article by Sir William Stubbs, who was Chairman of the QCA until September 2002; certainly not a casual observer of the system. Back in the olden-days, so the story goes, we used to define the A level standard in terms of norm-referencing. This meant that we awarded the same profile of grades across subjects, across boards and each year; regardless of how well the students had actually performed in each subject, each board and each year. Norm-referencing was, therefore, blind to the quality of work produced by students. Indeed, Sir Bill went so far as to describe this detachment from attainment as a source of invalidity, at least in relation to trends over time. The implication is that pass rate trends could not be interpreted as evidence of trends in national attainment over time; i.e. national attainment could be rising or falling but pass rates would still remain the same. The maintenance of A level standards, from a norm-referencing perspective, is straightforward: to apply the same standard, for any examination cohort, all you have to do is to apply the same percentage pass rate.

Although the myth of norm-referencing predates the 1960s, 1960 represents a very important chapter in this story. It saw the Third Report of the Secondary School Examinations Council, on A level examinations and the case for their reform (SSEC, 1960).

The A level was originally a pass/fail examination, certifying that students were qualified for university entry. However, by 1960, it had increasingly become an instrument for competitive selection. This meant that university selectors had started asking for numerical marks; which, in turn, had led to "an unhealthy competition in cramming and mark grubbing" by students (see SSEC, 1960, p.3). Moving from a pass/fail system to a graded system was supposed to remedy this.

The SSEC report proposed that there should be five passing grades, from A to E, and a compensatory O level pass. Although it did not actually specify how standards should be set or maintained, it did recommend that grades should be distributed roughly as described in Figure 1.

Grade	*Cum. %*	*Grade*	*Cum %.*
A	10	A	10
B	25	B	15
C	35	C	10
D	50	D	15
E	70	E	20
O pass	90	O pass	20
Fail	100	Fail	10

Figure 1: Recommendations from SSEC (1960)

The straightforward interpretation of these recommendations was as follows: irrespective of any possible difference in calibre of students between subjects, between boards or from year-to-year, the same percentage of students should be awarded each grade. That is, 70% should pass in German and 70% in Economics; 70% should pass with the Cambridge examining board and 70% with the London examining board; 70% should pass in 1960 and 70% in 1986.

Indeed, when looked at from a certain perspective, evidence does seem to suggest that this happened. A graph from the Centre for Education and Employment Research, at the University of Buckingham (BBC, 2010), nicely illustrated a striking norm-reference-like stability in the overall A level pass rate, from the early 1960s until the early 1980s (presenting data aggregated across subjects and across boards). From the early 1980s onwards, the pass rate rose steadily. An earlier report from the School Curriculum and Assessment Authority (SCAA) with the Office for Standards in Education (Ofsted) observed the same trend. A press release from SCAA (SCAA, 1996), which accompanied the report, read as follows:

> *From 1962 to 1986, the proportion of candidates to be awarded each grade in major A level subjects was effectively fixed, so no increase could take place even if candidates' performance improved. This was changed in 1987, when key grades were matched to the quality of candidates' work. This change from 'norm-referencing' to 'criterion-referencing' has permitted an increase in the proportion of candidates being awarded grades.* (SCAA, 1996)

This is the myth of norm-referencing: A level standards were norm-referenced, from the early 1960s until 1987, when they switched to being criterion-referenced.

The reality

It is straightforward to dispel the myth of norm-referencing, with reference to examinations results data that have entered the public domain every year since the A level came into existence.

Figure 2 represents data from a report published by the University of Cambridge Local Examinations Syndicate (UCLES, 1980). It illustrates differences in pass rates between syllabuses in different subject areas. The lowest pass rate was only 21%, which is clearly a vast distance from the supposedly recommended 70%. Admittedly, only 42 candidates were examined in accounting. Perhaps, then, only large-entry subjects were norm-referenced? The evidence suggests otherwise. The pass rate in

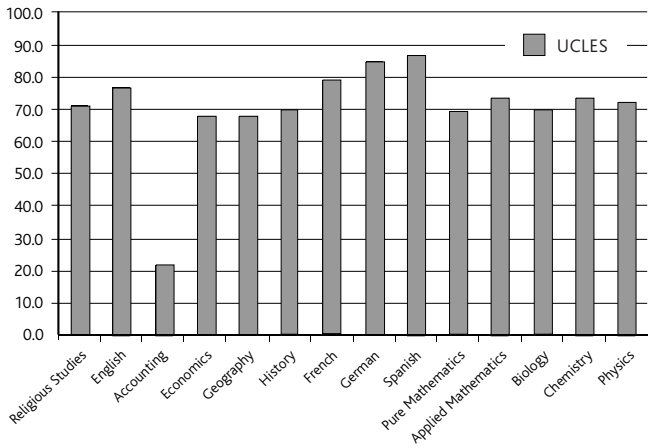

Figure 2: A level pass rate with UCLES, Summer 1980 (Home candidates only)

German was 85% (n=1336); in English literature, 77% (n=7519); and in economics, 68% (n=2699). This does not look much like norm-referencing.

Figure 3 adds data from the Associated Examining Board (AEB, 1980). It illustrates differences in pass rates between boards, within the same subject area. For geography, the pass rate was 68% for UCLES (n=4884) versus 48% for AEB (n=2247); for chemistry, the pass rate was 73% for UCLES (n=3288) versus 44% for AEB (n=2389); for French, the pass rate was 79% for UCLES (n=3335) versus 66% for AEB (n=2318). Moreover, the UCLES pass rates were almost universally higher than the AEB pass rates. Again, this does not look much like norm-referencing. Instead, it seems that the boards were quite clearly aiming to reflect differences in student calibre; both across subjects and across boards.

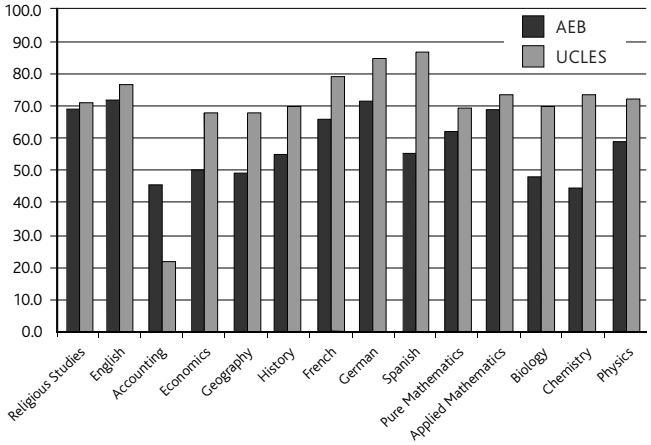

Figure 3: A level pass rate with AEB & UCLES, Summer 1980 (Home candidates only)

Maybe, though, the principle of norm-referencing was only applied within boards, within subjects, from one year to the next? Again, the evidence suggests otherwise. Figure 4 represents data from UCLES Annual Reports, from 1960 until 1986. Data from only four subjects were collated, for illustrative purposes. These are syllabus-level data – the level at which awarding decisions are made – so, if norm-referencing is to be found anywhere, it ought to be found in results like these.

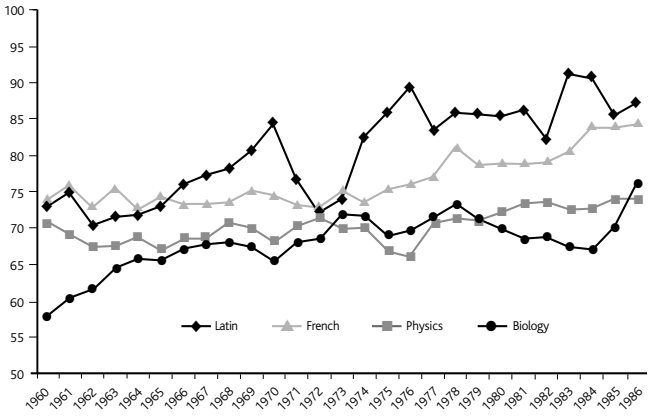

Figure 4: A level pass rates for UCLES (Summer, Home candidates only)

Even at this basic (non-aggregated) level the interpretation of the data is not entirely straightforward, particularly since syllabuses have a tendency to wax and wane in popularity and to be replaced. Where, for instance, there are two syllabuses in a subject area (with different pass rates), which are replaced the following year with a single syllabus (with a new pass rate), which of the two year 1 pass rates should the new year 2 pass rate be linked to? Of the four subjects in Figure 4, physics was the most tricky to interpret in this respect. For instance, in 1964 there was only one syllabus (physics), while in 1965 there were two (physics N and physics T); N gradually lost candidates while T gained them, until 1972, when there was only one syllabus again. To complicate matters, from 1974, physics split into physics and Nuffield physics. Similarly, 1974 saw biology split into Nuffield biology and social biology. Results are only presented for the highest entry 'conventional' syllabus (i.e. excluding Nuffield syllabuses and social biology). Results seem most straightforward to interpret for Latin and French, as these appeared to be essentially the same syllabuses from 1960 to 1986 (although, no doubt, they changed in content and emphasis over time).

Even for Latin, the entry sizes were reasonably large, from a low of 245 (in 1986) to a high of 939 (in 1964). Entries in the other subjects were higher still; in French, for example, ranging from a low of 1779 (in 1960) to a high of 3664 (in 1968). Given the proviso of large entry sizes, notice how:

- French jumped from 77% to 81% in one year (1977 to 1978)
- biology jumped from 70% to 76% in one year (1985 to 1986)
- Latin fell from 84% to 72% in two years (1970 to 1972).

Indeed, during the supposed glory-days of norm-referencing, from 1960 to 1986:

- the physics pass rate rose from 71% to 74% (+3%)
- the French pass rate rose from 74% to 89% (+15%)
- the biology pass rate rose from 58% to 74% (+16%)
- the Latin pass rate rose from 73% to 91% (+18%).

This was clearly not norm-referencing. Even though the pass rates do tend to rise somewhat less in certain subjects than in others, and even though there seems to be somewhat more stability in the 1960s than in the 1980s, it is still clearly impossible to claim that UCLES was norm-referencing at any point in time in any of these subjects.

Indeed, despite considerable research, I have uncovered no evidence that any board ever operated a principle of norm-referencing, although I have uncovered much evidence to the contrary. Choppin (1981) quoted Richard Christopher, Secretary of the Joint Matriculation Board, from 1977:

> *It is often thought that in pursuance of [constant standards] the percentages of candidates passing in a subject are decided in advance […] whereas the deciding factor is in fact the quality of the work presented.* (from Choppin, 1981, p.10)

Of course, when results are aggregated across syllabuses within subject areas, across subject areas and then across boards, year-to-year pass rate changes (for individual examinations) might well average out, giving the appearance of norm-referencing (at the national level). But this is not the same as examination standards actually being norm-referenced.

The confusion

If norm-referencing has never constituted a definition of the A level standard, then why does the myth persist? The answer is that something resembling it used to be used – and still is used – as a central component of grade awarding. I shall call it the Similar Cohort Adage.

The A level standard has always been defined in terms of a certain overall level of attainment, that is, the A level examination has always, essentially, been attainment-referenced. Indeed, this conception predates the A level and was a feature of the Higher School Certificate (from 1918) and of examinations which preceded that. However, equally engrained in the psyche of school examining in England is respect for the following adage: if the cohort hasn't changed much, then don't expect the pass rate to change much either.

As Christopher explained, the deciding factor in grade awarding has always been the quality of the work presented. But there has always been a healthy respect for commonsense and statistics too. So boards triangulate evidence from examiner judgement of scripts, with statistical expectations of pass rate stability, to decide where grade boundaries ought to lie. They did this in 1951, in 1960, in 1987 and we do it today.

If the principle of norm-referencing dictates 'any cohort – same pass rate' the Similar Cohort Adage recommends 'similar cohort – similar pass rate'. It is a rule-of-thumb that the examining boards in England have taken to heart and have integrated within their methodologies for maintaining standards. The Joint Matriculation Board set out its stall very clearly, in 1951, right at the outset of A level examining:

> *Many years ago in the light of its experience the Joint Matriculation Board reached the conclusion that the procedure fairest to the candidates was to award approximately the same percentages of passes, credits, very goods, etc. in every subject each year, the percentages being based upon the number of entries for the subject concerned. The Board insisted however that three strictly limiting conditions must all be satisfied.*
>
> *A. The subject entry must be large, many hundreds of candidates at least.*
>
> *B. The entries must come from a wide enough area to obviate the influence of special local conditions.*
>
> *C. There must be no reason to think that the general nature of the entries is changing.*
>
> (JMB, 1951)

The very fact that limiting conditions were identified illustrates that the JMB was not defining a principle of grade awarding, it was simply describing a rule-of-thumb to support grade awarding practice. The standard was defined in terms of student attainment; attainment-referencing.

Over the years, approaches to operationalising the Similar Cohort Adage have evolved. Early on, the boards had to rely on general impressions of whether, or how, cohorts were changing. To the extent that there were more boards in the 1960s, with smaller entries that were more locally based, this was more manageable. Fortunately, as entries have increased in size and boards have become less locally based, their statisticians have also become more sophisticated. With procedures like the delta analysis, they became better able to adjust statistical expectations, according to gender and school-type differentials (e.g. Eason, 1995). More recently, the boards have routinely established statistical expectations on the basis of prior attainment in the General Certificate of Secondary Examination (e.g. Pinot de Moira, 2008). Yet, conceptually speaking, the boards are not using statistical expectations any differently now from how they were used 50, or even 100, years ago.

The other myth

If the boards have always attainment-referenced and never norm-referenced then they can never have criterion-referenced either. Nor have they. This is the other myth. Criterion-referencing was certainly being considered, during the 1970s and 1980s, as a possible alternative approach to defining standards. Aspirations were particularly high, in some quarters, that the new 16+ examination (which was to become the General Certificate of Secondary Education) would be completely criterion-referenced. Keith Joseph proclaimed as much, in his 1984 North of England speech:

> *Second, we should move towards a greater degree of criterion-referencing in these examinations and away from norm-referencing.* (Joseph, 1984)

Yet, the GCSE was introduced with the traditional approach to grade awarding and with good justification for not criterion-referencing (see Cresswell, 1987; Cresswell & Houston, 1991). Despite high aspirations lingering for some time, criterion-referencing ultimately:

> *[…] died a death, other than taking the much weaker form of grade descriptions.* (Tattersall, 2007, p.70)

What actually happened in 1987

There is a grain of truth in the claim that norm-referencing came to an end in 1987. Only an extremely small grain, though. What happened was the result of a long campaign, spearheaded by the Joint Matriculation Board, to correct the narrow grade C problem; a problem that could be traced to the percentages recommended in 1960 by the SSEC.

Proposals for the reform of A level, within SSEC (1960), included the development of special (S) papers, based upon the same syllabus as A level papers, but examined at a higher level, giving abler candidates the opportunity to demonstrate their excellence. Discussions which preceded the publication of the 1960 report had concluded that S paper grades would only be awarded to students who had achieved at least grade B on their A level papers. Subsequent discussions led to the conclusion that it would be useful to lower this hurdle to grade C, just as long as the number of students who might additionally qualify was not too large. This is why the SSEC recommendations (in Figure 1) proposed a relatively narrow grade C band of 10% of candidates. As it happens, a ruling following the 1965 session lowered the hurdle to grade E. But the broad structure of the A level grade distribution had been established by then.

An unfortunate consequence of the narrow grade C was an increased likelihood of significant error in the award of grades. Students receive marks that fail to reflect their true level of attainment for all sorts of reason, from their own state of concentration on the day of the examination to errors made by clerical staff whilst inputting mark data. When grade boundaries lie only a small number of marks apart, the impact of this kind of error can be significant, for example, a genuine grade B student might end up being awarded grade D (or vice versa). The narrower the width between grade boundaries the more frequently these significant impacts will occur; an effect that is exacerbated when narrow grades fall at the middle of the mark distribution where the largest number of candidates often cluster.

In 1969, the JMB proposed an alternative grading system, based upon the following procedure:

1. first, set the top grade boundary;
2. then, set the passing grade boundary;
3. "All 'passing' candidates between the two fixed points would be placed on the agreed scale, by the application of a simple formula, strictly in accordance with the proportion of marks gained." (JMB, 1969, p.5)

In short, instead of grade distributions which had evolved on the back of a proportional division of *candidates*, the proposed approach would locate grade boundaries on the basis of a proportional division of *marks*. This would mitigate the problem of a narrow grade C. Unfortunately, after a concerted effort on behalf of the Schools Council, and major consultations with stakeholders, the Secretary of State dismissed the proposals in 1972. Of course, this did not resolve the problem.

A decade later, the JMB wrote another paper which, in effect, reminded readers of the problem of a narrow grade C and of the solution recommended in 1969 (JMB, 1983). This was debated for another three years, before a solution was finally agreed. The solution was not easy to reach. Although the similarities between boards in their approaches to maintaining standards far outweighed their differences, they were still somewhat precious over their differences and disagreed over how to resolve the problems of the A level grading system. A compendium of consultation responses received by the Secondary Examinations Council revealed comments such as the following:

> *The Committee accepted the need for a review and is not opposed, in principle, to a change in procedures for Advanced level grading. However, the present proposals are unacceptable to the Committee, primarily because of their effect on the level of difficulty of grades B and C but also because Chief Examiner judgements would come into play at only two grade boundaries*.
> (Letter from M.J. Jones, Examinations Secretary to the Welsh Joint Education Committee, to Sir Wilfred Cockcroft, 25 February 1985)

> *The Delegates point out that the SSEC guidelines which are at present used are no more than that: they are not rules to be rigidly followed, and while the effect of their application is one of the criteria studied by the Delegacy's Awarders in coming to their decisions, the final judgement is always made in the light of the actual work offered by the candidates. Flexibility is vital, in order to be free to do justice in subjects for which rigid rules are inappropriate, and also to avoid any unfairness to candidates which might arise from the use of an intractable framework. This is a matter on which the Delegacy's Awarders feel very strongly. The Delegates understand that there is to be no flexibility allowed in the use of the proposed system; this they deplore.*
> (Letter from C.G. Hunter, Secretary to the Delegates, University of Oxford Delegacy of Local Examinations, to Sir Wilfred Cockcroft, 4 March 1985)

> *The use of two fixed points in fact corresponds to this Board's practice over many years, where determination of A level grade boundaries commences with the establishment of what, in the judgement of the Awarders in consultation with the Secretaries, are the appropriate placings of the B/C and E/O boundaries*.
> (Letter from H.F King (Cambridge Secretary) and K. Schoenberger (Oxford Secretary), Oxford and Cambridge Schools Examinations Board, to Sir Wilfred Cockcroft, 29 March 1985)

> *Our views on which boundaries should be fixed by the quality of work shown in the scripts of candidates are divided. It could be that in the interests of public confidence in standards, the scrutiny of scripts at the A/B boundary is desirable. However the fixing of the B/C boundary in this way, involving as it does the inspection of the work of a larger sample of candidates, could produce a more 'reliable' performance indicator. On balance the weight of the argument seems to lie with determining the A/B boundary in this way rather than the B/C boundary.*
> (Letter from E.J. Bolton, Senior Chief Inspector, Department of Education and Science, to Sir Wilfred Cockcroft, 3 April 1985)

The University of London Schools Examining Board went a step further, arguing that changing the grading system in advance of the final outcome of research into the development of grade criteria was short-sighted. It recommended awaiting the outcome of a more fundamental review.

Ultimately, the solution to the problem of the squeezed middle was as follows:

> *There will be no change in the way grade A is awarded. The cut-off scores for grades B and E will be determined on the basis of the professional judgement of the examiners (and as this is, by and large, what is done at the moment there will be little change).*
>
> *Once the cut-scores for B and E are set, the minimum scores for D and C will be obtained by dividing the mark-scale into three equal sections*.
> (SEC, 1986)

Ironically, for some boards, the agreed procedure meant less reliance upon the judgement of awarders; not more reliance, as might have been assumed would be the outcome of the 'rejection of norm-referencing' (which, of course, it never was). For other boards, the 1986 ruling, to be operationalised from summer 1987, represented little more than business as usual. At the critical pass/fail boundary, at least, there was no change in procedure for maintaining standards, for any board.

A possible alternative explanation

Before concluding, an important alternative explanation for trends in pass rates over time needs to be considered; one that potentially resurrects the idea that examining boards were attempting to norm-reference after all.[iv]

Earlier, the principle of norm-referencing, in the context of UK examinations, was defined in terms of a requirement for awarding bodies to award the same percentage of students each grade in each examination. It was demonstrated that this principle was never observed since, at the syllabus level, different pass rates were evident across subjects, across boards and over time. What, though, if the principle were to be interpreted in terms of the *national* cohort of students attempting an examination in each subject area, rather than in terms of local cohorts of students attempting the same examination across individual boards? Perhaps the SSEC (1960) recommendations should be read as follows: 70% of students should pass each subject, at a national level, with candidate success spread across boards according to the calibre of students entered. For each subject, then, boards attracting more able students would pass more than 70% of students, and boards attracting less able students would pass fewer than 70% of students, such that the

national pass rate would average out to 70%. At first glance, this interpretation seems attractive. After all, there are clear indications that examining boards did indeed adjust pass rates to reflect the calibre of the entry for each examination. But, in doing so, were they ultimately aiming to norm-reference at a national level?

Against the national cohort interpretation of norm-referencing are a number of telling observations. For one, the boards themselves claimed not to be norm-referencing (see Choppin, 1981). Moreover, right from the outset of A level examining, it is possible to find very clear statements from the boards that they: (i) operated what I have called the Similar Cohort Adage at a local cohort level; but (ii) would not actually operate it if they had good reason to suppose that the general nature of an entry had changed (see JMB, 1951). In short, the boards were explicitly open to the possibility of pass rate change; even if, in practice, the national pass rate tended to remain fairly stable from the 1960s to the 1980s. Finally, if norm-referencing (*à la* national cohort interpretation) did indeed provide the explanation for pass rate stability until the 1980s, then who made the decision to stop applying the principle and for what reason? As explained earlier, the change in grade awarding procedure during the late 1980s – which was incorrectly described as the end of norm-referencing – did not significantly affect the way that passing grades were decided. It there had been an explicit policy decision to stop norm-referencing during the early- to mid-1980s, then it would surely have been well documented.

To evaluate the plausibility of the national cohort interpretation in more depth would require an analysis of data from all boards, going back to the 1960s and 1970s, to examine whether the cross-board, subject-level pass rates did indeed tend to average out to around 70%. These data, whilst potentially available, have not been collated for analysis to date. In the mean time, it seems more parsimonious to conclude that the local cohort interpretation of norm-referencing is both the conventional interpretation and a myth.

Conclusion

The idea that A level examination standards operated on a principle of norm-referencing until 1987, when they switched to a principle of criterion-referencing, is mythological but clearly false. In terms of the theory of grade awarding, 1987 saw:

- no rejection of norm-referencing as a principle (since it never has been assumed);
- no adoption of criterion-referencing as a principle (since it never has been assumed);
- no rejection of attainment-referencing as a principle (since it has always been assumed).

In terms of the practice of grade awarding, 1987 saw:

- no adoption of script comparison as a method (since examiner judgement has always been used);
- no rejection of the Similar Cohort Adage as a method (since statistical expectations have always been used).

Although the evidence which demonstrates this state of affairs is not always easy to locate, it is surprising that even official sources buy into the myth. One reason may be that the myth seems to provide a neat explanation for apparent changes in pass rates over time. At a national level, it is the case that pass rates have risen substantially since the 1980s; although, admittedly, they began their ascent during the earlier, rather than later, part of that decade. If A level awarding procedures did not change radically during the 1980s, especially not at the passing grade, then the pass rate trends are doubly remarkable. If we are to interpret the overall, national pass rate trend line at face value, then not only did student attainment rise substantially over the past three decades (during the 1980s, 1990s and 2000s), it rose from a baseline of no substantial change over the preceding two decades (during the 1960s and 1970s).

The only alternative explanation is that, despite the A level awarding process not having changed radically during the 1980s, more subtle changes were taking place, and these somehow affected the way in which grades were being awarded. This is certainly an intriguing possibility; but one beyond the scope of the present article.

References

AEB. (1980). *Statistics June 1980*. Hampshire: Associated Examining Board.

Baird, J., Cresswell, M.J. & Newton, P.E. (2000). Would the real gold standard please step forward? *Research Papers in Education*, **15**, 2, 213–229.

BBC. (2010). *A-levels: Rising grades and changing subjects*. 20 August. British Broadcasting Corporation Website. Accessed 24 March 2011. http://www.bbc.co.uk/news/education-11011564.

Choppin, B. (1981). Is education getting better? *British Educational Research Journal*, **7**, 1, 3–16.

Cresswell, M.J. (1987). Describing examination performance: grade criteria in public examinations. *Educational Studies*, **13**, 3, 247–265.

Cresswell, M.J. & Houston, J.G. (1991). Assessment of the national curriculum – some fundamental considerations. *Educational Review*, **43**, 1, 63–78.

Eason, S. (1995). *A review of the Delta Analysis method for comparing subject grade distributions across examining boards*. Guildford: Associated Examining Board.

Glaser, R. (1963/1994). Instructional technology and the measurement of learning outcomes: some questions. *Educational Measurement: Issues and Practice*, **13**, 4, 6–8.

Joseph, K. (1984). *Speech by the Rt Hon Sir Keith Joseph, Secretary of State for Education and Science*. The North of England Education Conference, Sheffield, 6 January.

JMB. (1951). *JMB Standardization and the GCE*. Manchester: Joint Matriculation Board.

JMB. (1969). *Advanced level grades: suggestions for a new system of reporting results*. Manchester: Joint Matriculation Board.

JMB. (1983). *Problems of the GCE Advanced level grading scheme*. Manchester: Joint Matriculation Board.

Newton, P.E. (2007). Contextualising the comparability of examination standards. In: P.E. Newton, J. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.). *Techniques for monitoring the comparability of examination standards*. 9–42. London: Qualifications and Curriculum Authority.

Newton, P.E. (2010a). Contrasting conceptions of comparability. *Research Papers in Education*, **25**, 3, 285–292.

Newton, P.E. (2010b). Thinking about linking. *Measurement: Interdisciplinary Research and Perspectives*, **8**, 1, 38–56.

Newton, P.E. (2010c). Conceptualizing comparability. *Measurement: Interdisciplinary Research and Perspectives*, **8**, 4, 172–179.

Pinot de Moira (2008). *Statistical predictions in award meetings: how confident should we be?* RPA_08_APM_RP_013. Guildford: AQA.

SCAA (1996). *Standards at GCSE and A level*. Press Release, 96/52. 5 December.

SEC. (1986). *SEC News. Number 3*. London: Secondary Examinations Council.

SSEC. (1960). *Examinations in Secondary Schools: Third Report of the Secondary Schools Examinations Council*. *The General Certificate of Education and Sixth Form Studies*. London: HMSO.

Stubbs, W. (2002). Gold standards and A-levels. *Education Today*, **52**, 4, 3–8.

Tattersall, K. (2007). A brief history of policies, practices and issues relating to comparability. In: P.E. Newton, J. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.). Techniques for monitoring the comparability of examination standards. 43–91. London: Qualifications and Curriculum Authority.

UCLES. (1980). *Statistics 1980*. Cambridge: University of Cambridge Local Examinations Syndicate.

Wikipedia. (2011). "Norm-referenced test", http://en.wikipedia.org/wiki/Norm-referenced_test, accessed 03 04 2011.

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

ENDNOTES

i. Of course, this raises the question of exactly how the pass standard for each sub-domain is defined and, likewise, how sub-domain standards beyond the passing grade are defined. The idea of criterion-referencing is most plausible for straightforward judgements (e.g. has vs. has not mastered) of low-level competences (e.g. the ability to add two single-digit numbers). In situations like this, the standard can be specified through fairly unambiguous written criteria (e.g. a student who demonstrates that they can add the full range of single-digit numbers, consistently over time and contexts, satisfies the criteria and can therefore be said to have mastered the ability to add two single-digit numbers). The more complex the judgement required (e.g. grade C standard vs. grade B standard) and the higher-level the competence in question (e.g. proficiency in statistics and probability) the less plausible criterion-referencing becomes. In situations like this, the standard is far less amenable to specification through unambiguous written criteria. Distinguishing the defining characteristics of competence at one level from competence at another becomes extremely complicated, as the competence becomes increasingly multifaceted and as students exhibit competence in a multiplicity of ways. Thus, the quest for unambiguous written criteria for the award of grades soon turns into a struggle to articulate even rough impressions. The rougher the impression conveyed by the written criteria the less meaningful the very idea of criterion-referencing becomes.

ii. The pros and cons of criterion-referencing and attainment-referencing are described in more depth in Baird, Cresswell & Newton (2000). The present article describes as attainment-referencing that which Baird, *et al* (2000) called weak-criterion-referencing. In retrospect, the term weak-criterion-referencing was not ideal. Attainment-referencing is conceptually quite distinct from criterion-referencing, not simply a weaker version.

iii. There are many ways of cutting the comparability cake; that is, many ways of expressing alternative conceptions of comparability (see Newton, 2010a). In terms of the tripartite framework presented in Newton (2010b,c), attainment-referencing and criterion-referencing would most likely be classified as examples of phenomenal definitions, defining comparability in terms of the observable phenomena of attainment (knowledge, skill and understanding). Norm-referencing, however, could not be classified within that framework (as either a phenomenal, causal or predictive definition). A fourth category of definitions would need to be introduced to accommodate it: competitive. Within norm-referencing, the standard is defined in much the same way as it is in any sporting competition: gold for first, silver for second, and so on. Results do not testify to an absolute standard ('the student has attained a, b or c'), merely to a relative one ('student 1 has attained more than student 2').

iv. Thanks to an anonymous reviewer for this suggestion.