INTER-SUBJECT STANDARDS: AN INSOLUBLE PROBLEM?

Ben Jones (Assessment & Qualifications Alliance), David Philips (New Zealand Qualifications Authority), Rob van Krieken (The Scottish Qualification Authority)

ABSTRACT

Whatever their constitution or the type of assessments they administer, it is a prime responsibility of all awarding bodies to engender public confidence in the standards of the qualifications they endorse, so that they have not only usefulness but credibility. Although guaranteeing comparability of standards between consecutive years is relatively straightforward, doing so between different subjects within the same qualification and with the same grading scheme is a far more complex issue. Whether standards are established judgementally or statistically – or, as in most contexts, a mixture of the two - satisfying public and practitioner opinion about equivalence is not easy. Common grade scales signify common achievement in diverse subjects, yet questions arise as to the meaning of that equivalence and how, if at all, it can it be demonstrated. With the increase in qualification and credit frameworks, diplomas and so forth, such questions become formalised through the equating of different subjects and qualifications, sometimes through a system of weightings.

This paper is based on two collaborative presentations made to the International Association for Educational Assessment (IAEA) conferences in 2003 and 2004. It summarises some recent concern about inter-subject standards in the English public examination system, and proceeds to describe three systems' use of similar statistical approaches to inform comparability of inter-subjects standards. The methods are variants on the subject pairs technique, a critique of which is provided in the form of a review of some of the relevant literature. It then describes New Zealand's new "standards-based" National Qualifications Framework, in which statistical approaches to standard setting, in particular its pairs analysis method, have been disregarded in favour of a strict criterion-referenced approach. The paper concludes with a consideration of the implicit assumptions underpinning the definitions of inter-subject comparability based on these approaches.

1. INTRODUCTION

It is a requirement for most awarding bodies to maintain comparable standards between different specifications and subjects as well as between years. The Qualifications and Curriculum Authority (QCA), the regulatory authority for public examinations in England, for example, specifies in its Code of Practice for awarding bodies that their "prime objectives are the maintenance of grade standards over time **and across different specifications within a qualification type"** (para. 110, emphasis added). The mechanisms for realising such equivalence are, however, often not made explicit by the regulatory or awarding bodies, although some have made attempts to address the question quantitatively and even qualitatively. For methodological reasons relating to the extensive assumptions which have to be made (discussed more fully later) statistical approaches to the issue have not yielded unambiguous, unproblematic conclusions. Qualitative approaches have been hampered by

the shortage of experts who are suitably qualified to make valid comparisons between disparate subject areas[1].

In recent years, however, the focus of concern, in England at least, has been less about comparability of standards over time and more about standards between subjects. The trigger for this concern has been the substantial changes in entry pattern in GCE A levels and the claims made by various newspapers and educational practitioners that these were largely driven by candidates opting for easier subjects. Headlines such as "Pass rates soar as pupils chase 'easy' A levels" (The Times), "Psychology, law and media studies: the 'scandalous' routes to A-level success?" and "Slump in languages as head teachers say pupils are opting for 'easy' A-levels" (both in The Independent) were, for example, common in the week of the GCE results publication in 2003.

It has been alleged that this in turn has been caused by a more mechanistic entry system to higher education, based on grade points regardless of subject, thus tempting students to opt for the "easier" subjects. Thus the leader of one of the U.K.'s leading teacher unions argues that:

> "Whereas admissions tutors used to give credit to students taking harder subjects … all subjects are now given equal credit in today's more mechanistic admissions system."
>
> Dunford (2003)

Dunford's arguments are supported by public examinations, and the grades achieved in them, becoming increasingly "high stakes" not just for the candidates who take them but also for their teachers and schools. Nor, is it argued, will the increasing establishment of qualification and credit frameworks, and collective "overarching" qualifications such as diplomas or baccalaureates, necessarily allay these concerns.

> "If you start to add up the marks on a baccalaureate-type award, and use overall totals to decide who gets in where, you will scare students away from packages that contain 'difficult' subjects, just as much as today. If you weight different subjects by their relative difficulty, then someone has to decide on, and justify, the precise weights."
>
> Wolf (2003)

Such concerns have not, however, merely been articulated in the pages of the national press. An independent panel of experts, invited by the QCA "to review the adequacy of the quality assurance systems that are designed to maintain GCE A level standards" showed concern in its report and commissioned a short-term research study by QCA[1], requiring it to:

> "Conduct qualitative analyses, in two subjects, of a series of examinations and resulting scripts detailing content and cognitive requirements. To judge comparability within and between subjects of the demands of the examinations and the standards of performance expected of students"
>
> QCA (2002)

[1] In England, the Qualifications and Curriculum Authority (QCA) is, however, currently undertaking a study of this type

Although this commission, which is currently being undertaken, is unusual in requiring qualitative comparisons between subjects to be made (most previous work in this area has adopted some kind of statistical equating), it is not a new departure by statutory bodies into this area.

While preparing his Working Group's Interim Report on 14-19 Curriculum and Qualifications Reform (2004), Tomlinson acknowledged in August 2003 that "there is evidence that some subjects appear to be harder than others", and Ken Boston, the Chief Executive of the QCA, argued that "geology, web design, cabinet-making and Latin [all] have value, but there is a need for comparability of standards between them" (Times Educational Supplement, 14 February 2003).

The following section describes similar methods from three examination systems which could be used statistically to align subject standards, and the degree to which each system used its method to inform its standard setting process. Each method is essentially a variant on what is often known as the subject pairs technique and embodies what Cresswell (1996) called the "same candidates" definition of standards. A critique of statistical methods for equating subject standards in general, and of these types of method in particular, is then provided in the form of a review of the relevant literature. The paper then describes New Zealand's new "standards-based" National Qualifications Framework, in which statistical approaches to standard setting, in particular its previously used "pairs analysis", have been disregarded in favour of a strict criterion-referenced approach.

2. THREE STATISTICAL APPROACHES

This section describes the statistical mechanisms which are used in three separate examination systems – Hong Kong, Scotland and England – to inform inter-subject comparability, and the extent to which the outcome of those mechanisms influence the award making process.

Hong Kong Examinations and Assessment Authority (HKEAA)

The Hong Kong examination system consciously aims to maintain comparability of standards in three dimensions – between consecutive years, between subjects and with other countries, particularly the U.K. The Certificate of Education Examination (HKCEE) is normally taken by students at the end of their five-year secondary education. There are forty four subjects which, with the exception of language related subjects, can be taken in either Chinese or English. Most candidates take seven or eight subjects, including English and Chinese. The same standards are applied in marking and grading, the language medium is not recorded on the results notices or certificates and results are expressed in terms of six pass grades, A - F.

The Advanced Level Examination (HKALE) is normally taken by students at the end of their two-year sixth-form courses. There are nineteen Advanced level and twenty Advanced Supplementary (AS) level subjects, with AS-level subjects being taught in half the number of periods required for A-level subjects, although to the same level of intellectual rigour. Most candidates take 5 subjects and, apart from Chinese Language & Culture and Use of English, which are again taken by most candidates, and other language-related subjects, all subjects can be taken in either language. The results are again expressed in terms of six pass grades A - F.

Although the focus of this paper is inter-subject standards, the following procedures describe the methodology used by the HKEAA to ensure comparability of its examination standards between consecutive years, between subjects and with other countries, particularly the U.K.

(a) Maintaining standards over time

The standards of most subjects are monitored year by year through a control group of schools (about one-third of the total) whose result averages fall within an acceptable range. Most control group schools have had stable results for five or more years, the stability being based on the stable academic ability profile of their intake. While the results of an individual control group school may be expected to show some variation between years, those of the group as a whole will not.

The HKEA adopts a norm-referencing approach to the grading of the main subjects. The percentages of control group candidates awarded the critical grades of A, C and E in a HKALE subject, for example, are reviewed annually and fixed in advance of the examination. (The percentages are proposed by the Grading Committee to the Board, and the Authority gives the final approval.) Once results statistics are available for each subject, the cut-off scores which will yield the predetermined percentages for the control group candidates are identified, and these scores are then applied to the whole candidature.

(b) Maintaining standards between subjects

The HKEA also tries to equate grade standards across all subjects at each level so that it is as difficult to gain a particular grade in one subject as in any other subject. This is done by defining the general ability of the candidates for a given subject in terms of their performance in all their other subjects. By 1992 all HKCEE major subjects, with the exception of three groups (details given in the paragraph below) were brought in line with this 'ability index' approach. In 1994, the methodology was revised to include the square of the correlation between subjects as a weighting factor in the determination of the ability index, so that subjects which correlated highly with the target subject had an increased weighting (or influence) on its ability index.

There are, however, three groups of exceptions to the ability index approach. The first group comprises skill-based subjects for which performance is judged by predetermined criteria, non O-level subjects, and subjects with fewer than 100 candidates. Subject comparability plays no part in the grading policy for these subjects.

The second group includes the two languages, Chinese and English, at both levels, and HKCEE Mathematics. These core subjects are taken by the majority of candidates and their grading has largely followed the historical pattern, adjusted by the results of the monitoring tests. The standards in English Language and Mathematics are closely linked to the standards of the University of London General Certificate of Education (GCE) Ordinary Level overseas examination, as part of the HKEA's arrangements to secure international recognition for the standard of its grade C awards in all subjects. This leads to low levels of award in English (about 8 per cent of school candidates gain grade C or better) and high levels in Mathematics (about 27 per cent gain grade C or better), reflecting the ability of Hong Kong students in these two subjects compared to that of candidates around the world taking the GCE Ordinary Level overseas examination.

The third group comprises HKCEE practical subjects. Before the HKEA was established, there were few such subjects, entries were small, and grading was left solely to the Chief Examiners. This led to a situation where, at grade E or better, awards were up to 30 per cent higher than in mainstream subjects. The HKEA has tried consistently to bring grading policies in these subjects closer to the ability index, causing some criticism to be levelled at its comparability procedures. Taking these concerns into account, the procedures were reviewed and modified in 1994, with the grading of the theory and practical components of these subjects now being separated. The grading of the practical component does not now involve the use of ability indices, and results of HKCEE practical subjects, along with those of HKCEE English Language, AS Use of English, AS Chinese Language & Culture, AS Liberal Studies and HKCEE Chinese Language, are reported in profile form.

(c) Maintaining standards with overseas examinations

The main way in which the HKEA used to attempt to ensure international comparability of its HKALE standards was by comparing the performance of students who took both an overseas examination and the local HKALE in the same year. On this basis, the results of eight HKALE subjects permitted comparisons with their U.K. A-level equivalents. These overseas comparisons used to show Hong Kong standards to be about one to one and a half grades higher for Applied and Pure Mathematics. At one stage HKALE Physics, Chemistry, Biology and Principles of Accounts used to suggest a higher standard than the U.K. A-level, although later they became on a par with each other. A possible explanation for this could be that latterly the UK examination placed more emphasis on communicative ability than previously, thus disadvantaging Hong Kong students. The HKALE Economics award also appeared to be of a higher standard than that of the U.K. A-level, but that in Geography and History did not, although the two systems' syllabuses for these subjects were very different, making comparisons difficult.

Currently, however, each year HKEAA sends three HKCEE and three HKALE examination papers and marking guides, together with a sample of marked scripts, to the University of Cambridge Local Examinations Syndicate (UCLES) for remarking and grading to enable benchmarking against comparable GCE Ordinary and A/AS-Level examinations. The reports provided by UCLES are used by the HKEAA Grading Committee to fine-tune its annual grading decisions. (Further information is available on the Hong Kong Examinations and Assessment Authority website, see references.)

The HKEA clearly faces conflicting priorities in ensuring comparability. On the one hand it goes to great lengths, using a stable common centres analysis, to ensure that standards are maintained in each subject (the large ones, at least) from year to year. As in most systems, this is the most important dimension of comparability as it would be unfair for a candidate's chance of success to depend upon the year in which she/he was born. On the other hand, the HKEA demonstrates an active concern that a candidates' chance of success should not depend on the subject(s) she/he selects and it also has an interest in trying to understand, if not maintain, parity of standards with other countries.

The Scottish Qualification Authority (SQA)

In 1999 the SQA introduced a new system of national qualifications which brought together into a single curriculum, assessment and certification system, subjects traditionally regarded as academic or general education and those traditionally perceived to be more vocational and work-related. They were brought together within the Scottish Credit and Qualifications framework.

Within this framework, a statistical approach to inter-subject comparability continues to be used for certain qualifications by the Scottish Qualifications Agency (SQA) to influence standard setting, as the following references indicate:

> "SQA also applies a system of national ratings which monitors performance between subjects and ensures that all subjects at the same level are broadly comparable in demand".
>
> From the SQA website

> "… paper B11/10 which related to the corporate goal 'to ensure that the award of all SQA qualifications is based on a consistent application of standards'. It was important to ensure that standards were maintained over subjects, levels, diets and years.
>
> In response to a question regarding comparison of respective levels of difficulty of current and new Highers, (it was) advised that initially use would be made of National Ratings which provide a statistical comparison. Specialist Groups would be asked to undertake checks on standards across levels."
>
> Minute from SQA board meeting, 26 September 2000

Each year the SQA publishes 'national ratings' for each subject offered for examination. These are comparability indices which inform the relative awarding standards in the various subjects at Standard Grade, Intermediate 1 and 2, and Higher and Advanced Higher levels, which are qualifications taken by students between four and six years of their secondary school education. The assumption underlying these indices is that candidates who, on average, do well in all subjects will also do well in any particular subject. While this assumption may not be true for a single candidate it may reasonably be applied to groups of candidates. The difference between a candidate's result in a given subject and the mean of the candidate's results in the other subjects taken is therefore, when averaged over a group of candidates, an indication of the "difficulty" of the subject in question. There is, however, a tendency for candidates to take groups of relatively easy or relatively demanding subjects, and to allow for this an adjustment is made.

A subject's national rating is the simple difference between the average grade performance in that subject, and the average performance in all other subjects taken by the same group of candidates, expressed in terms of grades. It thus shows how many grades higher or lower candidates obtained in the given subject than they did on average in their other subjects, with a positive rating indicating a relatively easy subject, and a negative rating a relatively difficult subject. For example, Standard Grade awards are expressed in terms of a seven-grade scale; a Standard Grade national rating of -0.50 would therefore mean that, on average, this subject's candidates were awarded grades of half a grade less than in the other subjects they attempted. All national ratings are derived from the results of candidates who attempted two or more subjects at a particular level. National ratings are not printed if the number of comparisons on which the calculation is based is less than twenty.

At Standard Grade the calculations are undertaken separately for all students and for the upper (above the median) and lower (below the median) cohorts, because the performance profile of examinations is different for candidates of different general attainment. For some

subjects, this can be half a grade or more in favour of either cohort. Male and female ratings are also calculated separately both for Higher and for Standard Grade. As with the upper and lower cohorts, the national ratings can be different for males and females, particularly at Standard Grade where the difference in some subjects can again be as much as half a grade. The national ratings also allow schools to produce their own relative ratings by calculating subject ratings within their school and subtracting the national ratings from them.

The national ratings have their limitations since there are several factors which are not taken into account, such as differences in the length of time for which candidates have studied a subject, differences in students' motivation between subjects and differences in the quality of teaching between subjects. National ratings of zero for all subjects might be considered desirable, but because of these, and other infringements of the underlying assumptions of the method, and the fact that there may be less need for direct comparability between academic and largely practical or creative subjects, outcomes from the national ratings are not applied strictly to align all subjects.

However, despite some limitations in the method, the Scottish education system pays more attention to these ratings than is normally paid to equivalent information in the rest of the U.K.'s system. There are three particular comparisons which can be made from the analyses to which SQA pays particular attention. First, because it is primarily concerned with maintaining continuity from year to year in the standard of each qualification, the national rating for each subject should not show large fluctuations, with the possible exception of subjects with small entry numbers. Second, cognate subjects in the same curricular mode with similar candidatures can reasonably be expected to have similar ratings. Third, comparing ratings of a subject at adjacent levels can assist articulation between the levels: for example a subject which appears easy at Intermediate 2 but difficult at Higher may cause problems for candidates. Finally, national ratings are also used by the Scottish Executive Education Department (SEED) to 'correct' individual schools' examination results, prior to distribution by local education authorities. Thus, although they are used to inform standard setting, national ratings are perhaps more assiduously employed in and by individual schools or local educational authorities to measure their performance in particular subjects against the national standard. Such localised ratings are called "relative ratings" and there is considerable advice about, and support for, undertaking this internal monitoring.

The English, Welsh and Northern Irish public examination system (EWNI)

Because historically there have been several awarding bodies serving the EWNI educational system, it is difficult to generalise about the ways they attempt to realise the requirement to align subject standards with each other. Historically, in England it was perhaps the Joint Matriculation Board (JMB, a predecessor of the current Assessment and Qualifications Alliance (AQA)) which used subject pairs analysis most widely as one of several indicators to inform awarding decisions. Following research based on the 1971 GCE Ordinary level examinations, Forrest and Vickerman (1982) reported that the JMB Examinations Committee maintained that subject pairs provided "an invaluable piece of additional information about comparative standards in examinations", and decreed that "analyses should be made of each year's examinations for the information of Subject Committees and examiners" (p.7).

Although subject pairs analyses can be designed in various ways, the method used by the JMB, as for most other awarding bodies that used the technique, was that implied by Nuttall, Backhouse and Willmott (1974). Candidates' grades were converted to numerical values (for

GCE A level, A=5, B=4 etc.)[2] and, for pairs of subjects with more than fifty candidates, the difference in mean grades calculated. To achieve an overall summary of the difference values for a given subject (i.e. its implicit relative severity or leniency), the values from each pairing were themselves averaged across all the pairings for that subject.

One of the possible variations on this method would be to weight a subject's pairings according to the correlation between the individual pairs (as in the HKEAA), or by the size of the pairing, when aggregating. It was, however, the simple, unweighted method which was used by the JMB and its successors for its subject pairs analysis for GCE A level, which is taken by most students after seven years of secondary education and results for which are reported against five pass grades, A – E. With the introduction in 1998 of the General Certificate of Secondary Education (GCSE), taken by most students at the end of five years of secondary education, it was recognised that a single difference value for a subject could mask variation across the longer A - G grade range. Consequently, whilst retaining the "mean grade method" for GCE, a "by grade method" was introduced for GCSE, whereby a separate subject pairs value for each grade was produced. A fuller explanation of this approach is provided by Fearnley (1998). This method also had the added benefit of treating the grade data more appropriately, i.e. as ordinal, not as equal interval in nature. One of the features of the JMB subject pairs analyses was the clear reporting of the assumptions underpinning, and caveats surrounding, the method. These were considered at length in Forrest and Vickerman (1982) and summarised in the annual subject pairs reports.

In 1993 the GCE boards collectively took the opportunity to use the emerging national, matched student-level datasets to commission their own research into grading standards, part of which was the execution of a national subject pairs analysis. In Chapter 5 of the report, Willmott (1995) briefly reviewed earlier analyses, before listing some of the factors which might legitimately explain apparent inter-subject differences in standard: teaching effects, assessment regime, the multi-dimensionality of achievement, gender effects, domain sampling, resourcing, motivation and interest, form of assessment, question difficulty and the distributions of marks. He consequently argued that, in the context of subject pairs in particular, terms such as 'leniency' and 'severity' should be used carefully, as difference values may not necessarily indicate a genuine difference in inter-subject standards. Although this work interested the GCE boards, its only effect was to make boards more wary of using subject pairs analyses in their awarding practices.

As a result of this work, and the compelling arguments against the validity of the implicit assumptions underpinning subject pairs techniques which are described in the following literature review, this type of approach to equating subject standards has played, at most, a minimal role in the awarding processes in each of the EWNI awarding bodies in recent years. Despite subject pairs analyses continuing to be routinely produced by some of them, they are at most only used as secondary information and are usually not referred to at all.

[2] Translating essentially ordinal data (candidate grade) into an equal interval level scale in this way, and analysing it accordingly (by means of calculation of mean scores) is, strictly speaking, inappropriate. Fowles (1996) described the issue in full, exemplified with some actual and some simulated data.

3. CRITIQUES OF STATISTICAL APPROACHES TO COMPARING INTER-SUBJECT STANDARDS

There are several papers which, from a theoretical perspective, undermine statistical methods for comparing subject standards, by reviewing the meaning of equivalence of inter-subject standards and questioning whether it could or should be achieved quantitatively. Goldstein & Cresswell (1996) and Newton (1997) questioned the assumptions upon which subject pairs analyses rest, Newton arguing that, to make statistical comparisons between subjects, one of two views had to be taken about the underlying ability measure(s) that the examinations were measuring: either there is a uni-dimensional underlying quality ("ability") common to all subjects, or there are multi-dimensional ability measures. He defined "ability" in this context to include all factors contributing to a candidate doing well in examinations, including motivation, parental support etc. and maintained that such factors had to be equal between subjects for the method to be valid. Newbould (1982) had argued that this was not the case by demonstrating empirically a relationship between apparent "ease" of subject and pupil preferences. Thus, apparently misaligned subjects may actually be in alignment if factors such as motivation were allowed for (although he did acknowledge that the argument could be reversed, i.e. that students tend to opt for "easier subjects" rather than perform well in those which motivate them). Newton also argued that justifying subject pairs analysis from the multi-dimensional position requires that the sample of candidates which takes a subject be representative of the total population of all possible candidates, which is unlikely.

The uni-dimensional position not only expects a substantial degree of correlation between subjects (because of general ability) but also that it is this ability which the various examinations are intending to measure. The issue of whether public examinations are intending to measure ability or attainment is a crucial one in terms of awarding in general, and comparability issues (including inter-subject comparability) in particular. The current awarding apparatus and procedures of the above systems imply that it is attainment that is being measured, albeit on the basis of "weak criterion referencing" (Baird, Cresswell and Newton, 2000), yet statistical approaches to comparability necessarily imply that an underlying measure of ability is being monitored. The summary of Newton's argument with subject pairs analysis is that it cannot legitimately equate attainment, as it cannot distinguish between differences due to subject specific factors (e.g. motivation) and those due to grading 'errors'. Tests can only be equated in terms of general ability or attainment; to the extent that the former varies between subjects, subject pairs analysis is unable to measure the latter.

However, in a more recent, unpublished paper, Newton (2003) suggested that the weak criterion referenced, attainment-based understanding of, and approach to, standard setting is relatively recent - in GCE at least - and that standards used to be, and perhaps should return to being, defined in terms of candidate ability. Thus, successive cohorts of students with identical ability profiles should have identical grade distributions, regardless of the extent to which, for example, changes in teacher motivation, teaching styles, learning resources and curriculum time were to affect the levels of student attainment. (If information about changes in the national levels of attainment was needed, it could be monitored via a sampling of students.)

This possibility raises the question of how to define and measure ability for linking standards, although this might not be as difficult as how attainment is defined and measured. In general terms he suggested that a common measure of prior achievement might be usable, although the operational details of the approach were less important than the theoretical consideration of what is meant by standards. By proposing linking standards to the ability of candidate cohorts over time, Newton also partially reopened the subject pairs debate. Whilst maintaining that "the ability definition presented above could support a coherent definition (of standards), as well as a

methodology based upon it (akin to the subject-pairs analysis)", he acknowledged that "if recommendations from such methodologies were fully implemented there would be few low grades in certain subjects and few high grades in others." Apart from the technical inadequacy of using the ability definition in this way, this situation would be politically unacceptable and might require an arbitrary inter-subject linkage to be made.

Goldstein & Cresswell (1996) also argued that "in a strict sense the assumption (of unidimensionality) is almost certainly false", both for subject pairs and reference test methods, and made two further criticisms of statistical approaches. First, they suggested that the degree to which the samples of candidates used in a subject pairing were typical of all the subjects' candidates should be reported. The method also implicitly discounts the effects of quality of teaching, general educational provision, students' interests, cognitive maturation effects and "the many other possible ways in which the quality of students' education in different subjects can differ" (p.9). Second, they rehearse the problems of inter-subject differences varying between identifiable sub-groups of candidates (e.g. defined by gender), whereby aligning the outcomes for the paired population could exacerbate the difference in outcomes for a sub-group. A related point is that the definition of standards implicit in subject pairs analysis is population dependent. Thus a change in, for example, the balance in gender in one subject would affect the relative difficulty of all paired subjects. They concluded that subject pairs analysis "cannot say anything in absolute terms about grading standards".

More recently, Wiliam (2002) conducted a review of standard setting in the national curriculum statutory assessments, where, because a pair of subjects is often taken by the population cohort, subject pairs analysis appears more valid. Cresswell (1996) refers to this possibility, arguing that if differential motivational effects are attributed to the learning characteristics of subjects then, for two subjects taken by the population, "the same-candidates definition of comparable standards (subject pairs analysis) is theoretically coherent and might be useful" (p.74). Even so, Wiliam recognised that even this 'norming' definition is weak since it fails to allow standards over time to be monitored because changes in the level of influential variables, e.g. motivation, are discounted. This is particularly relevant to National Curriculum assessments where the monitoring of standards over time is important. Wiliam was, therefore, dismissive of this approach for achieving any kind of understanding of inter-subject standards: "the question of whether standards of achievement in English are comparable to those in mathematics is not just difficult to answer in practice – it is a question that is fundamentally meaningless, except in the trivial sense that two norm-referenced tests are equally hard because the average scores on the tests are the same" (p.9).

Cresswell (1996) concurs with Wiliam in pointing out that statistically equating subject standards implicitly discounts features of the specifications (both content/demand and organisational aspects) which legitimately might affect candidate attainment. A commonly mentioned feature is the differential motivational effect between subjects, raising the question of whether one which stimulates candidates to perform better than in their paired subjects should be deemed lenient. Such features could, moreover, reasonably be expected to interact with others, such as differences in the demand of specifications, which ought to be equivalent, making it impossible unambiguously to equate standards statistically (see, for example, Jones, 1997). At the very least, it has to be assumed that such effects are equal across all specifications. Second, it was argued that the problem of different outcomes for identifiable sub-groups (see earlier) is compounded when they perform differently according to the techniques by which they are assessed. Finally, statistical equating requires that the specifications being compared are of equivalent demand, that is the value of what they are assessing is comparable. Thus, although it would be possible to conduct a subject pairs analysis across two levels, e.g. GCE and GCSE, equating the respective grade distributions would not produce equivalence of standards. Whereas equivalence of demand might be relatively easy to ensure when comparing

specifications within the same subject, making such judgements between disparate subjects is notoriously difficult.

Because inter-subject standards could not be defined in a straightforward statistical way, Cresswell preferred a value-based definition, derived from suitably qualified experts attributing to the standard their values of the attainment. Standards under this definition are less a reflection of the actual characteristics of attainment, and more an (expert) human response to that attainment, so there is no external statistical way formally to establish standards or indicate if they have been correctly equated. Historically this definition has, albeit implicitly, been adopted by awarding bodies in establishing subject standards and defending them against charges of misalignment. Incidentally, because it focuses on attainment it also conflicts with subject pairs analysis which necessarily assume that subjects are equated according to an underlying measure of ability (see Newton above).

Two empirical studies which cast doubt on the efficacy of statistical approaches to comparability are also worth noting. First, in 1996, the GCE boards commissioned a critical investigation into the validity of using statistical methods to evaluate inter-subject standards. Alton and Pearson (1996) compared subject outcomes by analysing the 1993 and 1994 EWNI national matched datasets according to the following methods:

- the prior attainment measure of GCSE performance;
- candidates' performance in pairs of subjects;
- candidates' performance in subject triples.

Despite some inconsistencies, all approaches tended to yield similar outcomes which reflected the pattern found in other studies (for example Nuttall *et al* (1974) and Forrest and Vickerman (1982)) and other countries (for example Elley and Livingstone (1972)) although Pollitt (1996) maintained that international similarities simply reflected a cultural correspondence (see later).

Arguably the most important outcome of Alton and Pearson (1996) was not the results of the analyses but information about the weaknesses of the methods. For example, when identifiable sub-groups of candidates (e.g. males and females or candidates from different centre types) were treated separately, the patterns of results, and the correction factors, showed some fairly large discrepancies. In addition, the implications of equating subjects according to their subject pairs value (or any of the other analyses) would have sudden, and sometimes substantial, effects on grade distributions (often for subjects with large entries) which could be both publicly unacceptable and educationally indefensible.

Second, in noting similar patterns occurring in the outcomes of statistical approaches in several countries, Pollitt (1996) suggested that this could be thought of as reflecting intrinsic differences in subject difficulty. He argued, however, that a more plausible explanation was that they reflected the existence of psychosocial phenomena which are only partially common across international boundaries. He illustrated this by comparing the apparent difficulty of GCE A level subjects with that of the same subjects from an unnamed Pacific Rim country, which also uses U.K. GCE A levels. Despite some similarities, Mathematics appeared to be relatively easy, and Business Studies relatively hard, the opposite of typical subject pairs values in the U.K. He suggested that factors related to students' motivation for choosing subjects would plausibly explain this outcome, concluding that "we cannot interpret differences between subject mean grades … as evidence of 'difficulty' unless we know who took each subject and why" (p.3). Deeper analysis of the data revealed further anomalies in relation to gender effects (e.g. the eastern males found history "easier" than mathematics; females found the opposite.) He concluded that "the only way to explain these oddities is by assuming that there are significant

differences between east and west in subject selection and hence in subject-specific ability and motivation. A subject pairs analysis is simplistic and dangerously misleading" (p.3).

4. NEW ZEALAND'S "STANDARDS-BASED" APPROACH

Since 2002, a new system of senior secondary qualifications has been progressively implemented in New Zealand as traditional examinations have been replaced by the standards-based National Certificate of Educational Achievement (NCEA). Starting in 2002, the School Certificate was replaced by the NCEA Level 1 (generally year 11 students, or 15-16 year olds); in 2003 the Sixth Form Certificate by the NCEA Level 2, and in 2004 University Bursaries (UB) by the NCEA Level 3 (generally year 13 students, or 17-18 year olds).

The old system

Prior to its demise in 2004, various procedures were used to ensure that students' achievement within and between UB subjects was normalised. Statistical scaling procedures were used to generate comparable medians and to determine the final distribution of students' marks for each subject. This was required because marks were aggregated for financial and status awards (e.g. the A and B bursaries, and scholarships) and for tertiary selection purposes, where ranking of candidates was required in order to allocate students to courses with restricted numbers. The maximum total scaled mark that a student could gain would be 500, although in practice totals higher than 460 were rare.

The procedure worked as follows. Where necessary, marks for school assessed components were moderated to ensure comparability between schools. This process was one of group adjustment, using the performance of the school group in the national examination (the common measure of all candidates) to adjust the school assessed marks. Raw mark distributions from the examination might have varied from one subject to another, for example a slightly easier or harder, or a longer or shorter examination, in comparison with others, could have resulted in the candidates in that subject being relatively advantaged or disadvantaged.

The moderated internally-assessed school marks and the examination marks were then assigned the specified prescription weightings and aggregated to obtain a total mark. Inter-subject scaling was carried out by adjusting each subject's standard scores such that the performance of its group of candidates was comparable to that of the group in their other subjects. The inter-subject scaling of marks was a percentile analysis process based on the 95th, 90th, 75th, 50th, 25th, 10th and 5th percentiles and applied to the national distribution of the marks for a subject. It was based on only those candidates who had entered three or more subjects who were not first language speakers in those subjects designated second language, thus helping ensure that candidates were not advantaged or disadvantaged by their choice of subjects.

Several assumptions underpinned this procedure, for example, that candidates' performance in any given subject is related to their performance in the other subjects in which they are being examined in the same year. This approach was based on norm-referencing in order to generate the final spread of marks for each subject. However, the types of knowledge and skills students were expected to demonstrate within each subject varied, and they were expected to chose different subjects depending upon their interests or future course preferences. Nevertheless, this method of scaling provided a definition of statistical comparability among subjects which aimed to ensure that candidates were not disadvantaged by the choice of subjects that they took.

The new system

Since the mid-1990s, the New Zealand Qualifications Authority (NZQA) has been implementing a new set of school qualifications within the National Qualifications Framework (NQF), of which the NCEA is part, which contains all nationally recognised qualifications, and is based on explicit standards. The development started around 1998, with a dedicated team in the Ministry of Education (the Qualifications Development Group) convening groups of subject experts and national standards bodies to develop and define achievement standards for all subjects that had traditionally been assessed through external examinations managed by NZQA.

Under this scheme, a 'standard' refers to a set of clearly defined and nationally recognised areas of knowledge, understanding and skills that a student must demonstrate to gain credits towards a particular component of an approved qualification. It thus defines the levels of achievement students need to attain in the various aspects of a subject in order to gain the credits that are attached to them. The standards include a mixture of internally assessed (i.e. as part of school-based coursework) and externally assessed standards (i.e. formal, terminal examinations administered by NZQA). Teachers use assessment schedules to help them judge whether the criteria have been met in the internal assessments and samples of these schedules and actual pieces of student work at each 'grade' from each school are externally moderated to ensure national consistency.

Achievement standards have been defined for traditional school-based academic subjects, previously assessed by School Certificate, Sixth Form Certificate and University Bursaries, and are registered for the NCEA in the NQF at levels 1, 2 and 3 respectively. They are generally based on the achievement objectives in the national curriculum statements that describe what students are expected to know, understand and can do. For each achievement standard, the required levels of knowledge, understanding and skill are specified, along with criteria for three levels of performance: Achieved, Merit or Excellent. In general, the achievement standards do not prescribe content or the full texture of a curriculum as in the national curriculum statements, nor do they prescribe how assessments are to be undertaken.

Unit standards are similar to achievement standards except that they do not discriminate between standards of performance according to the above levels. Students can gain credits if they meet the criteria specified in unit standards, of which over 16,000 have been registered on the NQF, covering virtually every area of education and training up to degree level. Credits from all of these unit and achievement standards can also count towards the NCEA. In order to obtain a NCEA level 3, a student would need to have 80 credits, of which 60 or more would need to be at level 3, and the remainder from any level.

As previously, Year 13 students will typically enrol on a course requiring study of four or five subjects. However, from 2004, their achievements have been assessed in terms of the criteria specified in the achievement standards. Students successfully meeting the assessment criteria specified for an achievement standard will gain the credits allocated to that achievement standard. A year's study of a subject is considered to be worth broadly the equivalent of 24 credits. A student following a full Year 13 course, with all subjects being studied at the same level (e.g. NCEA level 3), will therefore typically be assessed against achievement standards with an overall total of 120 credits. Many students will, however, achieve fewer than 120 credits either because they decide to study fewer than five subjects or because they do not meet the prescribed assessment criteria for one or more achievement standards. Other students will follow a course of multi-level study, with a mixture of achievement standards at NCEA levels 2 or 3, or even NCEA level 1, and they can also follow a variety of menus, including parts of subjects or extended study beyond 24 credits within a subject.

The main purpose of the new system is to assess student achievement against explicit standards rather than through a statistical ranking process. Through this, it was intended that more students could gain recognition for their achievements, particularly credits towards a wider range of nationally recognised qualifications, a broader range of achievements more closely related to the intended curriculum would be recognised, more differentiated information on student achievement would be provided to stakeholders, and more varied ways of assessing learning would be introduced. Consequently, students' achievement is now reported, not as a series of single subject marks (adjusted to equate statistically with each other), but as a profile of bands of achievement reflecting students' success in reaching both internally and externally assessed achievement standards, of which there are many in each subject.

The introduction of the standards-based NCEA has thus rendered the scaling of students' results between subjects or between years obsolete. This change in the nature of school qualifications aims to recognise and report in some detail what students have achieved in their courses, rather than merely position them in a rank order, which was the focus of the previous system. For the purposes of entry to Australian universities a new methodology has been developed to produce an index based on NCEA level 3 results that fairly represents New Zealand students' achievement.

5. DISCUSSION

This paper discusses the issue of what is meant by comparability of standards between different subjects within the same qualifications, and in particular whether it is amendable to a statistical solution. In surveying the types of statistical analyses produced by the Hong Kong, Scotland, England and the old New Zealand systems to inform this issue – all essentially based on a subject pairs methodology – and the differing uses they make of the results, several features emerge.

First, the method of grade awarding in each of the above systems is by holistical professional judgement on the quality of candidates' work, not on judging whether they have unambiguously met certain fixed criteria. This approach has been called "weak criterion referencing" (Baird et al, 2000), one of the critical features of which is that standards can be maintained through decisions about grade boundary marks being flexible to compensate for differences in the demand of the assessment across different dimensions, most especially time, and between subjects. As far as changes in demand over time are concerned these are manifested in even small unintentional variations in the demand of question papers and mark schemes. In the interests of fairness to candidates and public confidence in the system, most of the above systems value the maintenance of standards over time as paramount. Perhaps not coincidentally, ensuring of comparability of standards between consecutive years most easily lends itself to support by valid statistical information. One of the problems associated with comparing standards in this way is, as exemplified most strongly in the Hong Kong context, that there are several competing dimensions across which standards must be equated.

Perceived differences and discrepancies in demand and standard between subjects are, however, far more difficult to measure and compensate for, whether from a statistical or judgemental perspective. It would, of course, be possible to establish subject standards according to statistical procedures of the type described above. The definition of standards implicitly underpinning such an approach has been called the "same candidates definition" (Cresswell (1996)). According to this definition "two examinations have comparable standards if, when the same group of candidates is entered for them both, the distributions of grades which they produce are identical". Whilst appearing plausible, underpinning such an approach lie several assumptions which, as Alton and Pearson (1996) and others have demonstrated, can be very weak. For example, the definition assumes that "motivation, prior achievement and the influence of relevant school variables, together with the effects upon these of the two

syllabuses and examinations, are identical when the same candidates tackle two different syllabuses and examinations", assumptions which, as Cresswell (1996) argues, need not obtain even though the same students are involved. It is not difficult to imagine situations where, for example, the motivation of a group of students towards one subject is greater than towards another.

If the assumptions underpinning a statistical approach to inter-subject comparability are questionable, those required for a judgemental approach are no less insecure. Such an approach requires suitably qualified experts to judge levels of performance between diverse subjects - whilst compensating for different levels of demand in specification requirements, question papers and mark schemes - and pronounce about their equivalence. Identifying people competent enough to judge with authority the relative standard of attainment of diverse subjects, let alone make judgements which are publicly accepted, would prove difficult in most cases and impossible in some. It might, however be reasonable to assume that suitably qualified judges could be found who could compare cognate or semi-cognate subjects (e.g. between modern foreign languages), although even scrutineers in the U.K.'s same-subject inter-board comparability studies find it exacting and levels of agreement are usually low.

New Zealand's radical "standards-based" approach is devoid of any explicit statistical approach to the establishment of standards and making of awards, based as it is on a strong criterion-referenced design. In this design, standards are pre-determined by suitably qualified groups of subject experts, according to what they deem appropriate for prospective candidates to achieve. No possibility exists, therefore, for adjusting provisional awards post-hoc in order better to align standards either over time or between subjects, even if that were thought desirable. The approach does not, however, preclude subject pairs type analyses from being undertaken, but the standards are determined by fiat and, regardless of any statistical evidence to the contrary, they, and the awards they yield, are immutable. Whilst this may raise other issues related to standard setting - for example the inability to compensate for unintentional changes in assessment demand between years and the difficulties even expert awarders experience in making qualitative judgements with no quantitative supportive evidence (see, for example, Cresswell, (2000)) - it renders any debate of comparability of inter-subject and inter-year standards obsolete.

Although this strong criterion-based system seems substantially different from the weak-criterion referencing of the others discussed, they are in essence similar in their treatment of inter-subject standards. This may, however, be for different reasons which perhaps are more explicit in the New Zealand system than in the other systems. Public examination awards of whatever design must, above all, command public confidence which means they must be defensible. In the New Zealand system that confidence is embedded in the criteria which candidates are required to meet which have been pre-determined by subject experts. Although the system cannot demonstrate comparability of standards between subjects, in principle it can do so between years since the criteria against which they are set are fixed. Although the other systems cannot appeal so easily to such external evidence, confidence in them is embedded in a long history of grading which is transparent and yields awards which are known, understood (although there has been some disquiet about this in England at least over recent years) and accepted. The confidence in these systems is, therefore, based on their ability to demonstrate maintenance of standards over time, even during periods of major specification change and development.

Indeed it could be argued that the expectations of standards in each subject have become so embedded over time that even if there is a perception that some are harder than others then users (admissions tutors, employers etc.) and structures make tacit adjustments for that. Not only would such confidence be undermined by erratic changes to long-standing relationships between subject awards, caused by a subject pairs approach, but such changes would be

difficult to defend on the basis of the assumptions which underpin such methods. This view was espoused by Wolf, in an article discussing the desirable characteristics of "an English baccalaureate":

> "We currently maintain the polite fiction that all A levels are equivalent (and so a given grade gets the same UCAS points whatever the subject.) No-one believes this, but it doesn't matter because offers for most degrees are tied to specific grades in specific subjects."
>
> Wolf (2003)

In addition, in concluding that it may not be possible quantitatively to compare the standard of one subject with those of others, Newton also advises that either "we should learn to accept and adapt to the enigma of inter-subject comparability" or use the subjective value judgements of appropriately qualified experts.

6. CONCLUSION

Although some subjects' standards in any system may appear statistically misaligned, the assumptions underpinning both statistical and judgmental techniques for their alignment, not to mention the political reasons, are not compelling. In the absence of convincing theoretical, technical and political arguments that aligning standards between years is less defensible than between subjects, it thus appears appropriate that the former remain the focus of attention for awarding bodies, with the issue of inter-subject standards remaining essentially insoluble.

B E Jones
April 2005

References

Alton, A. and Pearson, S. (1996). *Statistical Approaches to Inter-Subject Comparability.* Unpublished UCLES research paper.

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

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

Cresswell, M. J. (1996). *Defining, setting and maintaining standards in curriculum-embedded examinations.* In Goldstein, H. and Lewis, T. *Assessment: problems, developments and statistical issues.* Chichester, Wiley.

Cresswell, M. J. (2000). 'The role of public examinations in defining and monitoring standards'. *Proceedings of the British Academy*, 102, 69-120

Dearing, R. *Review of Qualifications for 16-19 Year Olds: Full Report*. *Quality and Rigour in A level Examinations*. London, SCAA, March 1996.

Dunford, J. (2003). *We do still have to address the issue of equal standards for all A-level subjects.* Article in the Guardian Education, 26 August 2003.

Elley, W. B. and Livingstone, I. D. (1972). *External Examinations and Internal Assessments: Alternative Plans for Reform.* Wellington; New Zealand Council for Educational Research.

Fearnley, A. J. (1998). *Update on an investigation of methods of analysis of subject pairs by grade.* (Unpublished NEAB research paper.)

Fowles, D. E. (1996). *The translation of GCE and GCSE grades into numerical values.* (Unpublished NEAB research paper.)

Fitz-Gibbon, C. T. and Vincent, L. (1994). *Candidates' performance in public examinations in Mathematics and Science.* London; School Curriculum and Assessment Authority.

Forrest, G. M. and Vickerman, C. (1982). *Standards in GCE: Subject Pairs Comparisons, 1972-1980.* Occasional Publication No 39. Manchester; Joint Matriculation Board.

Goldstein, H. and Cresswell, M. J. (1996). *The comparability of different subjects in public examinations: a theoretical and practical critique.* Oxford Review of Education, Vol. 22, No.4, pp 435-442.

Hong Kong Examinations and Assessment Authority website address: http://eant01.hkeaa.edu.hk/hkea/new_look_home.asp

Janis, I (1972). *Victims of Groupthink: psychological study of foreign-policy decisions and fiascos* (2nd edition). Boston: Houghton Mifflin.

Jones, B. E. (2003). *Subject Pairs over time: a review of the evidence and the issues*. Unpublished AQA research paper RC/220.

Jones, B. E. (1997). *Comparing Examination Standards: is a purely statistical approach adequate?* Assessment in Education: Principles, Policy & Practice, 4 (2), 249-264.

Murphy R. J. L. *et al* (1995). *The Dynamics of grade awarding in GCSE.* School of Education, University of Nottingham.

Newbould, C. A. (1982). *Subject Preferences, Sex Differences and Comparability of Standards*. British Educational Research Journal. 8(2), 141-146.

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

Newton, P. E. (2003). *Contrasting definitions of comparability*. Unpublished paper.

Nuttall, D. L., Backhouse, J. K. and Willmott, A. S. (1974). *Comparability of standards between subjects.* Schools Council Examinations Bulletin 29. London, Evans/Methuen Educational.

Philips, D. (2003). *Lessons from New Zealand's National Qualifications Framework,* Journal of Education and Work, 16 (3), pp. 289 – 304.

Pollitt, A. (1996). *The "Difficulty" of A Level Subjects*. Unpublished UCLES research paper.

Qualifications and Curriculum Authority (2002). *Maintaining GCE A Level Standards: The findings of an independent panel of experts.* QCA. London.

Qualifications and Curriculum Authority (2003). *GCSE, GCSE in vocational subjects, GCE, VCE and GNVQ code of practice 2002/3.* QCA. London.

Stoner J. A. F. (1961). *A comparison of group and individual decisions involving risk.* Unpublished master's thesis, Massachusetts Institute of Technology, Boston.
Times Educational Supplement (14 February 2003) *Let industry mark exams.*

Tomlinson, M. (2004). 14-19 Curriculum and Qualifications Reform: Interim Report of the Working Group on 14-19 Reform.

Wiliam, D. (2002). *Level best? Levels of attainment in national curriculum assessment.* Report commissioned by the Association of Teachers and Lecturers (ATL), London.

Willmott A. S. (1995). *A National Study of Subject Grading Standards at A level – Summer 1993.* A report commissioned by the GCE Secretaries' Standing Research Advisory Committee.

Wolf, A. (2003). *An 'English Baccalaureate': exactly what do we want it to do?* Oxford Magazine, Eighth week, Trinity Term.