The Comparability of Different Subjects in Public Examinations: a theoretical and practical critique

HARVEY GOLDSTEIN & MICHAEL CRESSWELL

ABSTRACT *Comparability between different public examinations in the same and also different subjects, has been a continuing requirement in the UK. There is a current renewed interest in* between-subject *comparability, especially at A-level. The present paper examines the assumptions behind attempts to achieve comparability and explores the educational implications of some of the statistical procedures which have been advocated. Some implications for examination policy are also briefly discussed.*

INTRODUCTION

The public examination system in England and Wales has long been concerned with comparability problems. Comparability between different examinations in the same subject, set by different boards, has been a continuing requirement. The social requirement to maintain grade 'standards' over time for each subject has also, from time to time, been extended to include comparability between subjects at any given time. A summary of the various comparability procedures in use by examination boards can be found in Bardell *et al.* (1978).

Currently there is a renewed interest in between-subject comparability, especially at A-level where there is the associated problem of comparability with new vocational assessments such as GNVQ. The present paper examines some of the assumptions behind attempts to achieve comparability and explores the educational implications of attempting to ensure such comparability in an objective way. To illustrate our arguments we shall examine in detail a recent report which uses a particular approach to studying comparability (FitzGibbon & Vincent, 1994).

DEFINITIONS OF COMPARABILITY

An important distinction is between comparability of examination results and the equating of test scores. In the latter case (Holland & Rubin, 1982) the basic aim is to find a mathematical transformation which changes scores on test A, say, into scores on the scale of test B. If it is assumed that a common 'dimension' underlies both test score scales, this then allows individuals from a specified population who took test A to be assigned 'equivalent' scores on test B. In such procedures the assumption of 'unidimensionality' is crucial (Goldstein & Wood, 1989), and as we shall see, this assumption is intimately connected with certain kinds of comparability procedures.

In the case of examinations from different boards or in different subjects we have

0305-4985/96/040435–08

assessments which are linked to a syllabus or curriculum. Normal equating procedures, therefore, do not apply because the relevant populations for the two different examinations are distinct. They may be different in a number of ways, but most importantly they differ in that their curriculum and learning experiences are intentionally different, and this immediately poses a definitional problem.

Consider the case of two mathematics examinations at GCE A-level with papers set by two different boards, based on two different syllabuses. What *could* it mean to say that, for example, A grades from these two examinations were 'comparable'? If we were to ask real examiners, as is done in some comparability studies, the answer would be a variant of the following: that candidates awarded a grade A on either examination have demonstrated the same standard of mathematical attainment. Leaving aside problems associated with the reliability and bias of examiner judgements, the key point is that reference is made to a common 'standard of mathematical attainment'. This amounts to the assumption already mentioned of unidimensionality: that each syllabus and associated examination develops and assesses the same underlying attribute. In a strict sense, such an assumption is almost certainly false. Mathematics (and other subjects) are not homogeneous and different aspects will elicit differential responses among individuals.

Within a single subject area, there may be sufficient homogeneity so that, over the components of each syllabus, we can recognise a cluster of achievements which we are prepared to classify into a common set of categories. (One further assumption is also necessary: that the attainments of the candidates taking each examination have similar characteristics in terms of these categories. If they do not—e.g. those for examination A are better at most kinds of formal manipulations than those for examination B—this is evidence that the assumption of one underlying dimension is false.) When, however, the focus of attention is comparability between subjects, the need to assume unidimensionality is clearly a major impediment to its satisfactory definition.

In practice, however, explicit or implicit definitions of comparability involving the assumption of unidimensionality have usually been the ones operationalised either by 'cross moderation' (Bardell *et al.*, 1978) whereby examiners associated with each examination award grades on the papers from the other examination, or by using a 'reference test'. In the latter case, a 'common' test, say in mathematics, is given to a random sample of those candidates sitting both examinations and the statistical relationships between the scores on this test and the grades for the two examinations are used to adjust or 'align' the grades. We shall discuss the operation of such a procedure and the related 'subject pairs' procedure below.

There are, however, two alternative ways of defining comparability which avoid the unidimensionality assumption. One of these definitions is entirely statistical, the other is entirely qualitative. The first such definition, the statistical one, is as follows. Consider the situation where the students taking two different examinations are drawn (randomly) from the same population. This could occur, for example, if *all* students take the same mathematics and the same English examination. In other circumstances we might try to approximate such a selection by carrying out statistical adjustments for factors thought to be associated with the selection of students for particular examination courses. A consistent definition of comparability for two such examinations would then simply consist of allocating, as closely as possible, the same distribution of grades to each examination. We refer to this as the 'norming' definition. Apart from any practical difficulties for particular pairs of examinations, the principal difficulty with such a definition arises when we also require comparability over time. If the norming

definition is also adopted for comparability over time (as, for consistency, it must be) and if we therefore continue to allocate the same grade distribution at subsequent examinations in both subjects, we will be unable to observe the effects of any changes in the quality of teaching, for example, or in the population itself. The norming definition therefore has no utility if the examinations, in addition to providing information for selection of individuals, have any sort of monitoring function. In practice, of course, public examinations are sometimes required to perform monitoring functions, for example as the sources of data for school league tables, and in such a case it is difficult to see how the norming definition could be acceptable.

This brings us to the second, entirely qualitative, definition of comparability which avoids the questionable assumption of unidimensionality. The usual approach to maintaining within-subject-over-time comparability is to use examiner judgement to identify comparable performances in successive examinations. Similarly, examiner judgement can be made the basis for defining comparability between different examinations in the same subject and, indeed, between examinations in different subjects. If such judgements are accepted purely as qualitative value judgements which are made on behalf of society at large by people accepted as competent to make them, then there is no need to assume unidimensionality. This 'value-based' definition of comparability is discussed in detail in Cresswell (1996). Here, we shall simply note that, because value judgements are essentially subjective, the value-based definition avoids assuming unidimensionality only at the cost of explicitly not appealing to an objective description of examination standards. The extent to which such an explicit espousal of subjectivity would undermine the credibility of public examinations is unknown but clearly a potential problem.

Assuming that we are not prepared to forswear objectivity completely, if only for pragmatic reasons, we can see that, of the three remaining possible bases for defining comparability—equating, reference testing and population norming—none is fully compatible with the philosophy and purpose of public examinations. In the light of this we shall now look at one particular attempt to study between-subject comparability in an objective fashion and the assumptions, explicit and implicit, that it makes.

A LEVEL COMPARABILITY

FitzGibbon & Vincent (1994) (referred to hereafter as FV) analysed the A-level Information System (ALIS) database (FitzGibbon, 1992) and concluded that mathematics and science (MSc) subjects are more 'difficult' than non-mathematics or science (non-MSc) subjects. This follows previous work (reported by Tymms & Vincent, 1995) which concluded that at A-level, within each subject area, there was good agreement between boards and that high levels of comparability had been achieved.

The FV report uses essentially two main procedures. The first is based upon the group of students who take both MSc and non-MSc subjects and has two variants. One variant is the so called 'subject pairs' method whereby, say, those taking mathematics and French are used and if the mathematics grade, on average, is lower than the French grade then the mathematics is deemed to have been graded more severely. The second variant, a modification of this procedure, compares each subject with the average of all the other subjects taken by the student. This can be thought of as reducing 'sampling' errors by reducing the dependence of the results for the target subject of interest upon any single comparator subject. However, it does not raise any fundamental theoretical

issues which do not also apply to the more straightforward subject pairs analysis, and we shall not, therefore, consider it further.

FV's second main procedure is to use a 'reference test' measure on each student and then to compare, for each reference test value, average scores in each of the MSc and non-MSc subjects. The notion behind this is that resulting differences can be attributed to variations in the grading standards on the assumption that all other relevant differences have been allowed for. Of course, subject pairs analysis can also take account of a reference measure and the FV report partly recognises this by making some comparisons within three distinct reference measure groups. Two reference measures were used by FV: the average GCSE grade and a test taken in the same school year as A-level. We shall discuss the analysis based upon GCSE grades since this appears to be a better predictor of A-level results but our main reservations apply equally well to both reference measures.

SUBJECT PAIRS ANALYSIS

This procedure was studied extensively by the Schools Council (Nuttall *et al.*, 1974). The main technical difficulty with this approach, as the FV report points out, is that those students who happen to take particular pairs (or combinations) of subjects are not typical of either subject so that any conclusion is problematical. The report fails to quantify this so that the approach remains unconvincing. Nevertheless, for the sake of argument, suppose that everyone, or a representative sample, took, say, mathematics and French (this almost occurs with some subject pairs at GCSE). Suppose that, for such a group, the mathematics grades, on average, were lower than the French grades. On its own this would mean little for the following reasons.

We assume that the 'norming' definition is unacceptable for the reasons already discussed. This means that we would need in some way to satisfy ourselves that the students' pedagogical treatment had been 'comparable' in the two subjects before we could make any inferences about examination grading standards from their results. Quality of teaching and general educational provision in a subject influence examination results. There is a meagre discussion of the relative teaching quality issue by FV but, in any case, there are many other possible ways in which the quality of students' education in different subjects can differ. Students may develop interests, for example in foreign languages, which provide extra motivation for learning, or there may be some kind of cognitive maturational effect at work in some subjects more than in others.

Secondly, in practice we will find that although, on average, mathematics grades might be lower, there will nevertheless be many students for whom they are higher. Thus, even if we suppose, in some average sense, that mathematics grading is more severe we also have to admit that, for some students (perhaps even a substantial minority) mathematics is 'easier' than French. Also, the differences might vary systematically by, say, social class or ethnic group. If we were now to take action to make mathematics grades easier to achieve on the basis of the average result, in what sense would we have made the two subjects more comparable for the minority? The lack of a convincing answer to this question indicates the absence of any educational conception of examination standards underpinning FV's empirical work.

Our third point is related to the previous one. The definition of standards implicit in subject pairs analyses is wholly population dependent—it depends on averaging over the differences observed in an actual population of students. If, for some reason, the population characteristics were to change, say more girls took mathematics or there was

more exposure to foreign languages, then the relative difficulty of the subjects would change. Thus, it follows that this method cannot say anything in absolute terms about grading standards.

REFERENCE TEST PROCEDURES

The fundamental problem with reference measures is one we have already hinted at. Once it is admitted that some students find one subject most difficult but other students find another subject harder, we have accepted that performance is determined by at least two dimensions (if it were determined by only one then truly differential performance could not occur, apart from chance fluctuations). All the evidence suggests that achievement across subjects is multidimensional. Indeed, were it not, there would be little point in having different subject examinations or different curricula! Clearly, there is a variety of achievements, potentials, and so on which are differentially relevant to different subjects.

In such a situation, no single reference measure can allow appropriately for achievement in every subject. There will always remain an unexplained unique contribution within each subject. Moreover, this unique component will vary depending on the composition of the reference test. (There is, for example, no reason why average GCSE grade, as used by FV, with its different composition for each student, should be preferred as a reference measure over any other combination of achievements.) Since no reference measure can therefore adjust for *everything* which is relevant to examination performance *except* grading severity, we cannot use it as a basis for judging examination grading standards *per se*.

Procedures using reference measures involve treating residual deviations from a theoretical model as main effects when, in fact, the residuals must also be the result of all the other factors which have not been included in the model and whose effects are unknown. For example, as with the subject pairs analysis, reference measures cannot accommodate differential teaching quality or other subject-specific determinants of achievement. In addition, the results of reference measure analyses, like those from subject pairs analyses, are wholly population dependent (see also Wood, 1976, who provides an illustrative critique of between-subject comparability).

In the context of FV's GCSE analysis there are two further technical issues. In essence FV have averaged the subject differences over the range of GCSE scores. This is the outcome of their regression model. In practice, however, there are likely to be interactions, i.e. the subject differences may vary with GCSE score and also the relationship may be (indeed, almost certainly is) non-linear. Figure 1, taken from Goldstein & Thomas (1995) illustrates this for the relationship between total A-level score and GCSE.

The straight line is what FV would have fitted to the data and it can be seen how this leads to high scores and very low scores being under-predicted. If the relationship is non-linear and we fit a straight line averaging over the GCSE distribution, where we know that those taking some subjects, such as mathematics, do better than the average, this will bias the results. In addition, if, for example, there is less difference for the high GCSE performers than for the remainder (i.e. there are differential differences) this will further cause distortions when averaging over the whole GCSE distribution. Secondly, we note that FV have not done a multilevel analysis, so that their statistical tests will be wrong and further biases may well result (Aitkin & Longford, 1986, discuss this issue).

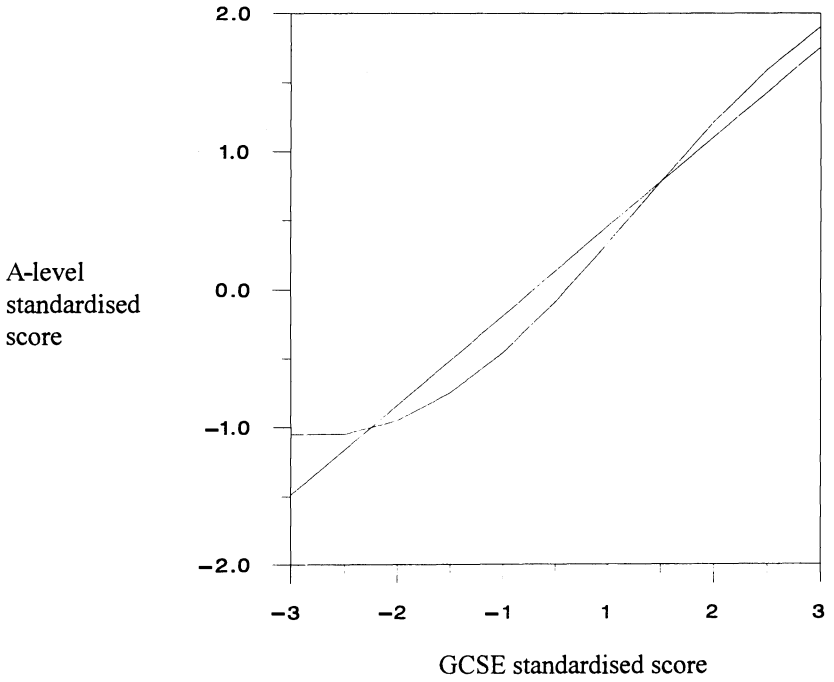

FIG. 1. Relationships between GCSE and A-level results, as modelled by FitzGibbon and Vincent (1994) (straight line) compared with empirical data reported by Goldstein and Thomas (1995) (curved line).

CONCLUSIONS

The fundamental problem with purely statistical approaches to the issue of comparable examination standards is that they ignore, as do FV, the educational content of the syllabuses and examinations concerned. Thus, it would be easy enough to set and grade a spelling test in such a way that it would be 'comparable' in FV's terms with A-level English; but would it be sensible, therefore, to accept the spelling test in place of A-level English as an entry qualification for degree courses in English? Alternatively, an examination could be set on a completely descriptive, non-mathematical physics syllabus and graded so as to be statistically comparable with current A-level physics examinations (or, indeed, art examinations). But would its syllabus be accepted by physicists as a valid or useful physics course for 16 to 18-year-olds? These hypothetical examples illustrate clearly the educational inadequacy of the definition of examination standards implicitly used by analyses such as that of FV.

For this reason, even if our more technical criticisms are dealt with, purely statistical procedures, such as those of FV, which rely upon the information available from a particular cohort of students, can never form a valid basis for judging, maintaining or adjusting examination grading standards. Moreover, in practice, the population dependence of such procedures, which was noted earlier, means that any changes in the composition and relative achievements of the population will destroy the possibility of maintaining comparability within subjects over time, insofar as this exists currently. Historically, within-subject between-year comparability has been paramount. However, our analysis shows why, even if this requirement were to be given up, both theoretical

and practical considerations would still preclude the possibility of making simple statistical adjustments to grade standards which would ensure between-subject comparability.

In the end, as always, the issue is one of policy. We are convinced, for the reasons we have given, that differences such as those reported by FV do not tell us anything useful about the comparability of examination standards. However, even if this is accepted in a theoretical sense, some may still ask whether, nonetheless, the existence of such differences poses some sort of practical problem for the public examination system and its users. Before attempting to answer this, it is important to anticipate the consequences which might follow from the use of results like those in the FV report to adjust examination standards.

If, as a matter of policy, a statistical definition of between-subject comparability (perhaps one of those used by FV) were to be adopted as a criterion, higher grades in A-level mathematics and science would presumably be made easier to achieve as a result. However, the consequential discontinuity in standards would raise such severe ethical and practical problems for selection between candidates examined in different years that it is difficult to see that it could be acceptable. Another consequence, if mathematics and science were to be made 'easier' in this way, is that more students would presumably be attracted to these subjects. At first sight, this might appear desirable, but the resulting 'grade inflation' would not necessarily produce a higher quality of mathematics or science student entering, say, higher education. It might also have a negative effect in the sense that it could be seen, and indeed be exploited, as a cheap method of 'improving' mathematics and science performance, rather than other approaches such as raising the quality of teaching, providing better resources, improving the morale of teachers and so on.

If these are some of the likely adverse consequences of adjusting public examination standards on the basis of analyses like those reported by FV, what would be the consequences of not adjusting them? The best answer to this question is provided by reflecting upon the fact that there is nothing new in FV's results. Many studies of different examination systems around the world have reported similar findings over many years (for example, Elley & Livingstone, 1972; Forrest & Vickerman, 1982; Nuttall, *et al.*, 1974, and many others). The consistent differences between subjects which emerge from such studies have not, apparently, prevented public examinations from meeting the perceived need for useful qualifications throughout this time. Moreover, some studies have demonstrated that the differences can be related, at least in part, to differential interest on the part of students (for example, Newbould, 1982). There is, therefore, no pressing practical reason to attempt to eradicate such differences by embarking upon a programme of adjustments to well-established standards, based upon theoretically invalid analyses which, in use, would create predictable, but new, practical problems.

REFERENCES

AITKIN, M. & LONGFORD, N. (1986) Statistical modelling in school effectiveness studies, *Journal of the Royal Statistical Society, A*, 149, pp. 1–43.

BARDELL, G.S., FORREST, G.M. & SHOESMITH, D.J. (1978) *Comparability in GCSE* (Manchester, Joint Matriculation Board).

CRESSWELL, M.J. (1996) Defining, setting and maintaining standards in curriculum-

embedded examinations, in: H. GOLDSTEIN & T. LEWIS, *Assessment: problems, developments and statistical issues* (Chichester, Wiley).

ELLEY, W.B. & LIVINGSTONE, I.D. (1972) *External Examinations and Internal Assessments: alternative plans for reform* (Wellington, New Zealand Council for Educational Research).

FITZGIBBON, C.T. (1992) School effects at A-level: genesis of an information system? in: D. REYNOLDS & P. CUTTANCE (Eds) *School Effectiveness, Research Policy and Practice* (London, Cassell).

FITZGIBBON, C.T. & 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, 1972–80*, occasional Publication No. 39 (Manchester, Joint Matriculation Board).

GOLDSTEIN, H. & THOMAS, S. (1995) Using examination results as indicators of school and college performance, *Journal of the Royal Statistical Society*, A, 159, pp. 149–163.

GOLDSTEIN, H. & WOOD, R. (1989) Five decades of item response modelling. *Brit. J. Math. and Statist. Psychol.* 42, pp. 139–67.

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

NEWBOULD, C.A. (1982) Subject preferences, sex differences and comparability of standards, *British Educational Research Journal*, 8, pp. 141–146.

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).

TYMMS, P.B. & VINCENT, L. (1995) *Comparing Examination Boards and Syllabuses at A-level: technical report* (Belfast, Northern Ireland Council for the Curriculum, Examinations and Assessment).

WOOD, R. (1976) Your chemistry equals my French, *Times Educational Supplement*, 30 July 1976.

Correspondence: Professor Harvey Goldstein, University of London Institute of Education, 20 Bedford Way, London WC1H 0AL, UK.

Editor's Note

Professor Harvey Goldstein has asked us to point out that on p. 430 of his joint article in the *Oxford Review of Education*, 19, 4 1993, 'A multilevel analysis of school examination results', the figure 0.09 in Table III and in the text below should read 0.29.