Contemporary Analysis in Education Series
Assessing Educational Achievement
Edited by Desmond L. Nuttall
The Falmer Press
(A member of the Taylor & Francis Group)

Models for Equating Test Scores and for Studying the Comparability of Public Examinations

Harvey Goldstein
University of London Institute of Education

It is often felt to be necessary, when different educational or other mental tests are given to individuals, to be able to 'equate' the scores on the different tests. Thus for two tests, for every score x on the one test a single 'equivalent' score y is needed on the other test. In this way we obtain a unique conversion, or transformation, from one scale to the other. It will then not matter which test is actually given to an individual, since all individuals can each be assigned a final score on the same scale. For example, if we wished to change tests over a period of time in order to avoid any one test becoming too widely known and thus easier for subsequent candidates, an equating procedure between the tests would still allow all candidates to be compared. Such a motivation lies behind the procedures adopted by the British public exam boards in their 'comparability' exercises.

It is possible to imagine a number of procedures for producing equivalent scores, for example, by transforming separately the distribution of each test score so that it has a standard normal distribution, which means that any score can be given the equivalent normalized score. Alternatively, a sample of individuals could each be given the two (or more) tests and a suitable empirical relationship between the test scores be used to transform one into another. In the following section some basic requirements needed for test scores to be equatable will be outlined, with a discussion of models which incorporate these requirements. Practical methods of estimating equating relationships will be referred to, and references to more detailed discussions will be given where they are available. A valuable

103

reference is the volume edited by Holland and Rubin (1982) which provides the most comprehensive account of modern test equating theory. The so-called comparability problem in public examination results will be discussed, and some suggestions will be made for alternative procedures.

The Theory of Test Equating

One of the fundamental assumptions in test score theory is that an individual's observed test score (X) consists of two components, his or her 'true score' (T) and a 'measurement error' (e) which add together to give the observed score thus

$$X = T + e \qquad (1)$$

where the mean value of e in repeated testing is $E(e) = 0$.

Typically, it is the true scores which we are interested in equating, although some authors have argued in favour of observed score equating (see, for example, Braun and Holland, 1982). A major problem with observed score equating, however, is that where measurement error distributions differ, the equated scores generally will have different population distributions, in particular different reliabilities. Thus, confidence intervals for true scores, or for the proportion of the population selected by a cut score, will depend on the 'parent' test — an undesirable feature.

In practice, once a true score equating function has been derived, it is the observed scores with which the function has to be used. This procedure may be justified on the grounds that an individual's observed score is an unbiased estimate of her or his true score and thus of the equated true score. In fact, other estimates may often be preferred. For example, if estimates of reliability and other population distribution parameters are available, so-called 'shrunken' estimates of true scores could be used. A practical difficulty with such procedures would arise, however, if individuals are able to choose which estimates to use, for example, by choosing to be regarded as a member of one particular sub-population so that their equated score could be maximized. The advantage of the observed score is that it is an unbiased estimate for any given set of circumstances. Potthoff (1982) gives a detailed discussion of methods based on the equating of true scores estimated from observed scores.

Suppose that we have two tests with observed scores X, Y, whose true scores S, T are to be equated. For equating to be possible we require every score S to be equivalent to one and only one score T in a strictly increasing or decreasing order. Thus, in the population, every individual with a given true score, say S, on the first test will have an equated true score, say T_1, on the second test.

We can write this formally as:

$$S_1 \equiv T_1 \tag{2}$$

If we now consider the observed scores X, Y then (2) becomes

$$E(X|S_1) = E(Y|T_1) \tag{3}$$

where $E(X|S_1)$ stands for the mean value of the observed X for an individual with true value S_1; likewise for $E(Y|T_1)$. Equation (3) is referred to as the 'weak' definition of equating. Lord (1977 and 1980) proposes a strong definition of equating. Not only does he require (3) to be true but also, after equating to a common scale, that the distribution of X about S_1 is identical to the distribution of Y about T_1, in particular that the variances of the corresponding measurement errors are equal. Lord justifies this additional requirement on the grounds of 'equity' by which he means that an individual who is equally happy whether he takes test 1 or test 2 must, rationally, want the measurement accuracy of each test to be equal, arguing that a test with a small measurement error variance should be preferred to one with a large measurement error variance. This assumes, however, that individuals have a particular kind of 'utility function' with a very high 'cost' attached to having an observed score a long way from the actual true score. Alternative utility functions are quite plausible, however. For example, large measurement errors will be associated with large over-estimates as well as large under-estimates and an individual, particularly one with low ability, may well prefer to 'gamble' on turning up a large over-estimate of ability. Lord's condition therefore seems to be too constraining. It is also very restrictive in effectively limiting the types of test which can be equated to those which are strictly parallel. In fact, in the later discussion (Lord, 1980), he is forced to consider practical methods of approximate equating for the majority of tests which do not satisfy his extra condition. It seems more sensible and realistic, therefore, to avoid that difficulty and to take equation (3) as the fundamental definition of equated tests (see also Morris, 1982).

We now need to specify how to operationalize expression (2), that is, to define a 'transformation' of the S scores to the T scores. For example, it might be a simple linear transformation

$$T = a + bS$$

and in general we may write T as a function of S

$$T = f(S) \tag{4}$$

where f(S) defines a monotonic relationship, that is, a relationship that is one to one and preserves the ordering.

Equation (4) can be extended readily to a series of tests. Such a series of related tests forms a 'unidimensional' set in the sense that once an individual is assigned a true score on one test, his true scores on the others are also uniquely defined. Note, however, that each separate test itself need not be composed of a unidimensional set of items, so that the test scores might, for example, be determined by a combination of two or more factors.

While (4) may refer to any monotonic relationship, it is simplest to begin with a linear one. A suitable 'model' for this case is the one known as the congeneric test score model described by Jöreskog (1971), the simplest version of which is

$$X_i = a_i + b_i T + e_i \tag{5}$$

where i refers to a test, X the observed score on that test, T the true score, e_i a measurement error and a_i, b_i scaling or equating parameters.

The usual assumptions for this model are

$$\text{covariance } (T, e_i) = E(e_i) = 0$$

and for convenience we can set

$$E(T) = 0, \text{ variance } (T) = 1$$

If $a_i = 0$ and $b_i = 1$ then the tests are known as tau-equivalent and if in addition the variances of the e_i are all equal then the tests are parallel.

The problem of equating then becomes the one of finding good estimates of a_i, b_i for each test, since when these are available, if we define

$$X_i' = (X_i - a_i)/b_i \tag{6}$$

then we have for two tests i, j

$$E(X_i'|T) = E(X_j'|T) = T \tag{7}$$

which is simply equation (3) with T being the true score on the common single dimension. Thus (7) satisfies our definition of equated scores and the transformation in equation (6) is known as a linear equating procedure (LP). Note that (7) does not require the measurement error variances to be equal so that we do not require tests to be parallel.

Now the variance of X is b^2_i/R_i and the mean of X_i is a_i, where R_i is the reliability of the *i*th test. Thus if we have a good estimate of R_i then we can estimate b_i and a_i by $\{R_i \text{ variance } (X_i)\}^{1/2}$ and X_i respectively. Where several tests are to be equated, efficient 'maximum likelihood' methods are available (see, for example, Werts *et al.*, 1980).

While the linear model (5) is relatively easy to deal with, in practice many relationships are non-linear. In principle (5) could be extended to include non-linear terms, but this would not only complicate the analysis, but it would also be difficult in any one case to know precisely which non-linear terms to include. A more flexible approach is the so-called equipercentile (EP) procedure. The aim of this is to rank in order the true scores on each test in order to obtain the cumulative probability distributions and then equate the equivalent percentile values. If a general non-linear monotonic relationship given by (4) exists, then since the whole population of individuals will be ranked (on their true scores) in exactly the same order by each test, an equating of the percentiles of the cumulative probability functions of the true scores will produce the required result. As with the LP method we do not require equal measurement error variances, but we must take care in the estimation. This is because the mean value of a percentile estimated consistently from an observed score distribution is not equal to the same percentile of the true score distribution. From equation (1) we obtain the usual relationship

$$\text{variance (X)} = \text{variance (T)} + \text{variance(e)}$$

with

$$\text{R} = \text{variance(T)/variance(X)}$$

Thus, a given percentile, say the ninety-fifth, corresponds to different values of the observed and true score distributions, and the observed scores need to be 'shrunk' to correspond to the distribution of true scores. If we assume that the distributions can be described in terms of their means and variances then we simply need to multiply the observed values (measured about the mean) by the square root of the reliability. It is then the percentiles of these shrunken distributions which are equated. Of course, when the measurement error variances are equal, then the raw scores can be equated directly. In order to obtain good 'smoothed' estimates of the cumulative distributions a combination of 'eye-fitting' and automatic procedures such as spline-fitting will usually suffice, although large samples will be necessary in order accurately to locate the extreme percentiles.

More recently, latent trait models have been used for equating. Accounts of the procedure can be found in Marco *et al.* (1980) and Peterson *et al.* (1982). Briefly, these models relate the responses of the constituent items of a test to an assumed unidimensional 'ability' for each individual and to one or more parameters relating to each item. Using either separate random samples or common tests, the item parameters for all tests can be estimated, thus enabling the ability of an individual who responds to any of the tests to be estimated. The use of latent trait models has been advocated for 'vertical test equating' where the performance of groups of markedly different ability is to be equated. Because they deal with items rather than the test scores, latent trait models require special assumptions, such as unidimensionality of items, to be made. These, however, give rise to difficulties (Goldstein, 1980) and the use of such models seems problematical. Moreover, Peterson *et al.* (1982) find that not only latent trait but also traditional equating procedures perform badly when the tests to be equated differ in difficulty, which is the case in vertical equating. Morris (1982) demonstrates theoretically that tests containing different ranges of difficulty cannot be equated even weakly and also shows that in general, if tests are multidimensional, equating procedures based on total test scores cannot be expected to work and it will be necessary to equate sets of component one-dimensional subtests where each set depends on a single dimension. It is possible for two tests each to be multidimensional and still be equatable in terms of total scores, but only in the special case where they have the same set of one-dimensional components whose loadings on the total score are related by a single linear function.

Educational attainment tests typically are multidimensional and so in general can be equated only via component one-dimensional subtests. This implies that such components have to be identified or, in other words, knowledge of the dimensionality structure of the test is needed. This is no easy matter and seems to have been little attempted in the context of test equating. Alternatively, if in order to make tests easily equatable, they are constructed to be one-dimensional then this carries implications for their validity. As population changes occur, either through curriculum innovation or for other reasons, so any attempt to maintain a sequence of one-dimensional tests begins to look very problematical (Goldstein, 1983).

Reference Populations

The previous discussion has referred to the equating of tests for a given population. The empirical equating literature, however, tends to be a little vague about the appropriate reference population for any given procedure. For example, that a procedure for equating two tests works well in one population does not guarantee it will do so in another. Furthermore, an equating procedure can work satisfactorily in a population but poorly in a sub-population — for example, a minority ethnic group. There is then an urgent question of the circumstances under which identifiable sub-populations 'gain' or 'suffer'. This is an empirical issue which has hardly been addressed at all by existing studies. Thus, the most common justification for the use of equating methods seems to be the existence of high (dis-attenuated) intercorrelations between the tests used. Part of the high intercorrelations, however, may well be explained by other factors such as socio-economic group, income, curriculum, etc., so that 'partial' correlations within relatively homogeneous sub-groups may be much smaller.

Equatability

Much of the equating literature seems to take the view that two tests either are equatable, or they are not. Since, short of studying every member of a population exhaustively, we cannot ascertain whether a procedure is perfect, we need some measure of how good a procedure is. Traditionally, a linear correlation coefficient is used, but this seems inappropriate for equipercentile methods. A rank correlation would be better and the following suggestion provides such a measure and suggests how it can be used to improve the practical application of an equating procedure.

Consider a population of individuals who take test A and test B and assume true scores are available, although in practice observed scores will be used. Then we can define perfect equating such that

$$X_{ia} > X_{ja} \Rightarrow X_{ib} > X_{jb} \qquad (8)$$

where i, j refer to individuals and X to test scores. If we have n individuals arrange the X_{ia} in ascending order and for each of the ${}^{n}C_{2}$ pairs of X_{ia} scores see whether (8) holds. The proportion of pairs for which it does is our index E. Clearly, if E is near to 1.0 we may be content with our procedures. If not, then we may be able to improve

matters by amalgamating or grouping scores on both tests to eliminate cases where (8) does not hold. Of course, this will not always be possible, but one would reasonably expect most inconsistencies to arise from nearby scores so that grouping these will lead to a higher value of E. Algorithms to do this could be programmed readily. Thus, we would produce a hierarchy of E values, from the original minimum value upwards. For any E we would have modified test score scales and when one was reached which was thought 'acceptable' this would give the corresponding procedure for equating the grouped test scores. We can regard the functions which carry each original score to the grouped scales as an expression of the loss of score precision required in order to carry out an 'acceptable' equating.

If E is used to measure equatability then we would want to report this for as many sub-populations as possible before recommending that the procedure is used in the total population. Thus, a high value of E over the whole population might be considerably reduced within sub-populations if the variable defining the sub-population was strongly associated with the test scores.

Designs for Equating

The first systematic attempt to devise a framework for equating studies seems to have been that of Angoff (1971). He proposed four main designs, and the following summary is based on these, incorporating the models of the previous section. (The case of just two tests is used for illustration.)

1. Each test is given to a different sample of the population. For the LP Method, equation (6) is used to equate to a common scale with a_i, b_i estimated using the reliabilities and means as given in the previous section. For the EP Method the 'shrinking' procedure is used separately for each sample.
2. Each test is given to all individuals in a sample, with the administration in one order for a random half and the reverse order for the remainder. This uses individuals more efficiently (by cutting down on the numbers needed) and the 'crossover' design enables allowance to be made for possible practice effects. Angoff's method, while incorporating an adjustment for practice; does not make explicit use of the relationship between the tests, although this can be incor-

porated in the congeneric model (5) to obtain improved estimates. In the non-linear case, efficient EP methods are complicated but estimates based on the separate distributions can be used. The relationship information does, however, allow a check on some assumptions.[1]

3 An additional common test U is given to each group in design (1). The purpose is to increase precision by adjusting for sampling fluctuations in the selection of the groups, using 'regression estimation' procedures for the LP method. Any variable with a fairly high correlation with the scores can be used for U, or indeed a combination of variables can be used. For the EP method, assuming a large enough sample, an iterative non-parametric standardization procedure can be used. Details are given in Bianchini and Loret (1974).

4 A common test U is administered as in (3) but U is now used to predict the true scores, with scores predicted by the same value of U deemed to be equated. Alternatively the tests may be used to predict U, with scores predicting the same value of U deemed equated. These methods seem not to be justified by any general model, but are used in public examination comparability exercises and they will be discussed more fully in a later section.

In evaluating the performance of these designs it is useful to assess how closely the sample data conform to the model. For this purpose we can define the conditional variance of equating (D) as follows:

if $X'_1 = X'_2$ then for test 2

$$D_2 = E\{(X'_{2j} - X'_2)^2 | X'_{1j} = X'_1\}$$

is the variance of the second test score values about the equated score for all individuals (j) with the same first test score.

Empirical Studies

The most comprehensive equating study so far has been the Anchor Test Study, commissioned by the US Office of Education and carried out by the Educational Testing Service from 1971 to 1974.

One part of the study, which is not of prime concern here, was a norming study involving 150,000 children. The test equating part of the study involved a stratified random sample of 200,000 fourth, fifth

and sixth grade children from the whole of the US and seven tests (with one added later in a supplementary study). One of the tests, the Metropolitan Reading Test, was chosen as the 'Anchor' Test (and was the one which was normed) and the others were equated to the scale and norms for this.

The study design consisted of sixteen replications of a basic design involving twenty-eight schools each given a testing assignment at random. For the seven tests there are twenty-one possible pairs and each test had a parallel form giving another seven pairs. Then within each school the testing was repeated using the reverse order to that first assigned. This resulted in 2 × 28 = 56 ordered pairs of tests. The final report is in thirty volumes and describes the results, and a project report (Bianchini and Loret, 1974) of 295 pages gives details of the design and methodology of analysis. Both LP and EP methods were used to obtain equated results.

Several studies have compared latent trait models with LP and EP methods, for example, Holmes (1980), Marco *et al.* (1980) and Peterson *et al.* (1982), but no one method emerges as clearly superior, and few useful simulation experiments seem to have been attempted.

The Comparability of Public Examinations

The General Certificate of Education boards in England, Wales and Northern Ireland issue graded certificates to individuals for each examination subject. Each board issues grades A, B, C, etc., in a particular 'O' level subject, carrying the implication that a grade A from one board is 'equivalent' to a grade A from any other board. As with test equating, therefore, an implicit equivalence relationship underlies the award of grades. I will begin by describing briefly how two common methods of equivalencing operate and then consider what theoretical underpinnings these may have. A more detailed description of the methods can be found in Bardell *et al.* (1978).

Monitor or Reference Tests

In this method, for each examination paper to be equivalenced, the examination score or, more usually, grade (using a simple scoring system) is regressed on a 'reference' test score. The difference between the intercepts of the regression lines (assuming them to be parallel) estimates the differences in the mean grade scores. These difference scores can then be used as the basis of adjustments to grade definitions in order to equivalence the mean grades with respect to the reference test. A detailed description of the workings of this procedure with examples can be found in Newbould and Massey (1979).

Apart from any theoretical difficulties, several practical difficulties occur with this procedure. Firstly, it may not be possible to adjust grade boundaries to produce coincident regression lines, and this will be so particularly if the original lines are not parallel or show signs of non-linearity. Secondly, the use of a simple scoring system for the grades is rather crude. Although it seems not to have been tried, a direct method of relating proportions of candidates in each grade to the reference test score would be preferable, using, for example, a logit linear model. Thirdly, some account should be taken of the measurement error in the reference test; it appears that only one research study has attempted to do this (Willmott, 1977).

Cross-Moderation

This has now become the favoured method and since 1978 all nine GCE boards have taken part in cross-moderation exercises at 'O' (Ordinary) and 'A' (Advanced) level.

Subject experts (usually examiners) scrutinize examination scripts to decide whether grades are 'comparable' across boards. This is done either by using a wide range of scripts from each board in order to establish where grade boundaries should be, or by using narrow ranges of scripts centred on grade boundaries determined by each board *a priori*. In the latter case it is often found that examiners from one board find another board too lenient, whereas the other board's examiners find the first board too lenient! This indicates that each examiner is using his or her own criteria, based on particular examination experience, to make judgments. To overcome this, attempts have been made, often involving outside experts, to evolve common criteria for these exercises. Nevertheless, agreement on criteria is not easy, and the result may be a compromise which is not as relevant to any single board as were the original criteria.

The advantage claimed for cross-moderation is that it comes close to the actual examining process, allowing the full use of expert judgment. On the other hand, it tends to be costly so that in practice only relatively small samples of scripts can be compared. It is also,

ultimately, subjective and dependent on which examiners or experts are used.

Both the reference-test and cross-moderation methods may be used either to compare different boards in the same subject in one year or to compare different examinations in the same subject for a single board for two or more years. The first application is designed to ensure that every candidate is treated 'fairly' or 'comparably' irrespective of which board's examination is chosen, and the second is designed to ensure that examination 'standards' remain constant over time. The reference test method has also occasionally been used to study comparability between subjects, but in the light of the following discussion this seems especially difficult to justify.

In the previous paragraph words such as 'fairly' and 'standards' have been used somewhat imprecisely, and little attempt has been made to provide a strong justification for the methods, unlike those underlying equating. In the next section I will attempt to outline the logic of a comparability model for public examinations, and then to see whether the procedures used actually satisfy the requirements of the model.

Models for Comparability

Perhaps the simplest procedure which could be used to attempt to obtain grade or score comparability would be a direct application of equipercentile equating using the cumulative grade or score distribution of each examination. An obvious objection to this is that the students taking the various examinations cannot be viewed as random samples from the same population. To overcome this, both the reference test and cross-moderation procedures attempt to 'adjust' for such student differences. Thus, the reference test is assumed to be a measure of student ability which captures such differences. The difficulty is that there is no simple ability or attainment which can be measured objectively (in the reference test case) or subjectively (in the case of cross moderation). The very point of having different syllabuses is to promote different abilities and attainments. Hence, as well as being different in degree, student attainments are different in kind since different aspects of a subject will have been studied and learnt, corresponding to the different exam syllabuses. Such deliberate diversity precludes representation by a single score on a reference test or by the average judgment of a set of examiners. The argument may be formalized in the following way.

For a given examination subject, consider two boards, A, B, and two syllabuses 1, 2. Syllabus 1 is the appropriate one for board A's examination and syllabus 2 for board B's examination. That is to say, each examination is designed to test attainment in the subject as described in the appropriate syllabus. Of course, in practice there are several boards and often more syllabuses than boards, but this raises no new issues of principle.

Now consider a hypothetical experiment whereby half of the candidates following syllabus 1 are allocated at random to paper A and the other half to paper B, and likewise for syllabus 2. For those candidates from syllabus 1 we compute the mean score difference between paper A and B, say x, and likewise for syllabus 2, say y. Since the allocations are at random, the average ability of the candidates is the same for each examination, so that we have the possibility of using the differences x, y for each syllabus separately, as adjustments to the examination marks, so that on average we can be fair to all candidates irrespective of which exam they take.

Unfortunately, since each examination is linked to a syllabus, we would expect those from syllabus 1 to do less than justice to themselves when taking examination B and vice versa for syllabus 2 examinees so leading to different values of x, y. Thus any supposed difference in examination difficulty is confounded with the examination/syllabus link and indeed x and y may even have opposite signs. In effect, this underlies the apparent contradiction found in cross-moderation exercises mentioned previously. In addition there is the practical difficulty that nominally the same syllabuses in different institutions may, in reality, differ considerably in the emphasis given to various topics, hence making them effectively different syllabuses. Moreover, this hypothetical experiment involves random assignment of candidates which is usually quite impractical. Nevertheless, the above argument will apply to other methods of adjusting for ability differences, such as reference-test and cross-moderation methods. The former uses an objective regression or covariance model to judge which candidates are equivalent, that is, have the same ability, and the latter method judges which candidates are equivalent according to subjective criteria developed by one or more moderators, this time using the internal evidence from the examination answers themselves. For both methods the average score difference for equivalent candidates is used to adjust examination scores. We see, therefore, that there can be little theoretical justification for the usual between-board comparability exercises.

Nevertheless, there is one special case when it would be

appropriate to attempt to adjust for 'ability', namely where for a single examination board there are equally relevant examinations for a syllabus. This might apply over time where comparability was desired from one year to the next. Here, however, there are additional problems related to the fact that syllabuses could change from year to year so that the relevance of a reference test to the examinees may change, as might the moderators' criteria.

Having shown that the current attempts at comparability have no adequate theoretical justification, it is relevant to ask whether an alternative theoretical model exists. Imagine, again, a hypothetical experiment in which individuals are initially randomly assigned to one or other syllabus. This would give, on average, equal distribution of ability at the outset, and if it were possible to ensure equality of education provision, teaching, etc., then if both groups take the same examination, any difference in score distributions would reflect differential relevance of the examination to the syllabuses, apart from sampling fluctuations. If there are now two different examinations, each related to one syllabus, then the difference in scores will reflect both 'relevance' and 'difficulty'. Nevertheless, it could be deemed fair in this case to use this difference to adjust scores, since the two groups of students are assumed to be equivalent. This imaginary experiment does seem to be the strongest sense in which public examination comparability can achieve fairness but, as before, we need to ask how closely the hypothetical experiment can be approached.

Firstly, neither the cross-moderation nor reference-test methods come close, since both rely on assessing examinees at the end of exposure to a syllabus. In principle, it would be possible to attempt to measure 'abilities' prior to syllabus allocation and also factors associated with teaching, etc. In practice no comparability studies along these lines seem to have been carried out, and to do so would involve a time-consuming longitudinal study. In addition to the above factors, moreover, variables such as student choice would have to be measured, since generally the choice of which examination to take is not made at random. In practice we know relatively little about how to measure the relevant factors associated with teaching or student choice. While further research aimed at understanding these is worthwhile, clearly we are far from possessing the knowledge needed to create satisfactory comparability exercises.

It should be noted that the above arguments are not limited to current, largely norm-referenced methods of examining. They apply with equal force to attempts to produce so called 'criterion-referenced' examinations. Even were such attempts successful in producing examinations with any notable advantages, they would not provide inherently a solution to comparability, and could make such problems even more intractable (see Orr and Nuttall, 1983).

Some Conclusions and Recommendations

This review has been generally critical and pessimistic about the utility of the various equating and comparability methods in use. It has been my intention to try to illuminate the logical foundations of these methods, in order then to evaluate the procedures themselves. In equating, there seems to be a need for some realistic simulations to evaluate the performance of different methods on data with known properties. It is worth pointing out that current interest in the provision of 'graded tests' in English and Welsh schools seems to imply large scale equating procedures in order to establish a working system (Nuttall and Goldstein, 1984.) In comparability, some long-term studies would be useful, but simulations of the conditions of student choice, examination choice, etc., would also be useful.

Where there are several examinations related to a single syllabus, it is possible to make progress towards establishing comparability, or at least deciding what degree of comparability might be attainable. This would seem to be the case with certain examining bodies such as the Business and Technician Education Council, where common syllabuses are separately examined by different institutions, and Nuttall and Armitage (1984) investigate models which make allowance for various student characteristics, including previous attainment. They show that it is possible to use their procedures as screening devices to identify potentially aberrant examinations, so that a more detailed study of these can be undertaken.

If reasonable comparability is not possible, perhaps we should be asking whether attempts to achieve it should not be abandoned. Why not, for example, have simple norm-referencing, whereby every year each set of examination scores is separately standardized using those individuals entering for it, and a common grading system used? This would at least have the merit of being well understood. Objectors to such a system might argue, for example, that this would penalize those children who happened to encounter a particularly 'difficult' paper, but it could also be said that any 'unfairness' introduced by this would be small in comparison to other known sources of variation, such as marking variability. It is also possible that after such a system had been in operation for several years, both those who

take examinations and those who use the results might accept the system fairly readily. The students would make their own decisions about their prospects with different examination boards, and the users would make allowances for different 'standards' adopted by the boards. Naturally, the boards would wish to maintain stable 'standards' but those would be incorporated into the setting of the examination papers. Since these papers themselves and the objectives of the syllabuses upon which they are based would be publicly available, the onus for a valid interpretation of the examination results would rest with the user rather than the present somewhat shaky comparability procedures. Furthermore, in those cases where valid exercises might still be carried out, such as overtime for a single board with an unchanging syllabus, these would provide a useful check on examination standards.

Acknowledgement

This chapter has benefited from helpful comments by Dr. J. Houston, Dr. M. Cresswell and Professor D.L. Nuttall to whom I am most grateful.

Note

1 In order to satisfy equation (4) a further assumption is necessary, namely that $E\{f(X_i)|S\} = f(S)$, with a similar condition for the other tests. However, this ought to be the case so long as the reliabilities are not too low. Also, this assumption can be examined empirically.

References

ANGOFF, W.H. (1971) 'Scales, norms and equivalent scores' in THORNDIKE, R.L. (Ed.) *Educational Measurement*, Washington, D.C., American Council on Education (2nd edn).

BARDELL, G.S., FORREST, G.M. and SHOESMITH, D.J. (1978) *Comparability in GCE: A Review of the Boards' Studies, 1964–1977*, Manchester, Joint Matriculation Board.

BIANCHINI, J.C. and LORET, P.G. (1974) *Anchor Test Study, Final Report: Project Report*, Berkeley, Calif., Educational Testing Service.

BRAUN, H.G. and HOLLAND, P.W.L. (1982) 'Observed-score test equating: A mathematical analysis of some ETS equating procedures' in HOLLAND, P.W. and RUBIN, D.B. (Eds.) *Test Equating*, New York, Academic Press.

GOLDSTEIN, H. (1980) 'Dimensionality, bias, independence and measurement scale problems in latent trait score models', *British Journal of Mathematical and Statistical Psychology*, 33, pp. 234–46.

GOLDSTEIN, H. (1983) 'Measuring changes in educational attainment over time: Problems and possibilities', *Journal of Educational Measurement*, 20, pp. 369–78.

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

HOLMES, S.E. (1980) *ESEA Title 1 Link Project, Final Report*, Salem, Oregon State Dept. of Education.

JÖRESKOG, K.G. (1971) 'Statistical analysis of sets of congeneric tests', *Psychometrika*, 36, pp. 109–33.

LORD, F.M. (1977) 'Practical applications of item characteristic curve theory', *Journal of Educational Measurement*, 14, pp. 117–38.

LORD, F.M. (1980) *Applications of Item Response Theory to Practical Testing Problems*, Hillsdale, N.J., Lawrence Erlbaum Associates.

MARCO, G.L., PETERSEN, N.S. and STEWART, E.E. (1980) 'A test of the adequacy of curvilinear score equating methods' in WEISS, D.J. (Ed.) *Proceedings of the 1979 Computerized Adaptive Testing Conference*, Dept. of Psychology, University of Minnesota.

MORRIS, C.N. (1982) 'On the foundations of test equating' in HOLLAND, P.W. and RUBIN, D.B. (Eds.) (1982) *Test Equating*, New York, Academic Press.

NEWBOULD, C.A. and MASSEY, A.J. (1979) *Comparability Using a Common Element*, Cambridge, Test Development and Research Unit (mimeo).

NUTTALL, D.L. and ARMITAGE, P. (1984) *A Feasibility Study of a Moderating Instrument*, Report to Business and Technician Education Council.

NUTTALL, D.L. and GOLDSTEIN, H. (1984) 'Profiles and graded tests: The technical issues' in *Profiles in Action*, London, Further Education Unit.

ORR, L. and NUTTALL, D.L. (1983) *Determining Standards in the Proposed Single System of Examining at 16+*, London, Schools Council.

PETERSEN, N.S., MARCO, G.L. and STEWART, E.E. (1982) 'A test of the adequacy of linear score equating models' in HOLLAND, P.W. and RUBIN, D.B. (Eds.) *Test Equating*, New York, Academic Press.

POTTHOFF, R. (1982) 'Some issues in test equating' in HOLLAND, P.W. and RUBIN, D.B. (Eds.) *Test Equating*, New York, Academic Press.

WERTS, C.E., GRANDY, J. and SCHUBAKER, W.H. (1980) 'A confirmatory approach to calibrating congeneric measures', *Multivariate Behavioural Research*, 15, pp. 109–22.

WILLMOTT, A.S. (1977) *CSE and GCE Grading Standards: The 1973 Comparability Study*, London, Macmillan Education.