

The Measurement of Standards

DAVID J. BARTHOLOMEW

The Concept of a Standard

STANDARDS ARE AT THE HEART of current educational debates and, if we examine the contexts in which the term is used, we find that the concept is essentially quantitative. Standards are spoken of as high or low and institutions and individuals are ranked according to the level of achievement. Yet, in spite of the ubiquity of the term, there is no simple measuring instrument we can take to a child or a school and read off the appropriate value. Neither is there any natural unit of measurement as there is with some physical quantities. Indeed, the more closely we analyse the concept, the more elusive it seems to be. We find ourselves in the position of St Augustine who is reported to have said of time 'What is time? If no one asks I know, but if I have to say what it is to one who asks, I know not'. Much the same could be said of standards.

The situation we have described is not uncommon in the social sciences; indeed it could be argued that it is the norm. Many of the key variables which occur in the discourse of the social sciences have the characteristic that they cannot be directly observed. Conservatism, alienation, attitude to abortion are further examples of things which are spoken of as if they were real quantities, yet they are not open to direct observation. The abilities and skills with which educational measurement is concerned are of just this kind. In the language of statistics they are latent variables because they are thought of as underlying, and influencing, the world of observation but are not, themselves, directly accessible to observation.

Before we can begin to provide a conceptual framework for

approaching the measurement of standards, we need to note that the term *standard* is used in a variety of senses and these need to be disentangled if confusion is to be avoided. All of them have something to do with levels of achievement, or performance, by children, students or institutions. The standard reached by a student at the end of a course, for example, is thought of as a point on a scale of performance which can be compared with that of other students or with some threshold for passing an examination or achieving a degree of a particular class. Here we are speaking of an *individual* scale which measures the performance of a person.

We also use the term in a collective sense to apply to *groups* of individuals, as when we use it of a school. School league tables are intended to rank schools according to their institutional achievements. Usually this will be done by aggregating the individual performances of the members of the institution by quoting, for example, the proportion who achieve certain A-level grades. These may then be combined with other, similar, measures to form a composite index. But in essence, institutional measures are derived from those pertaining to their members.

Thus whether we are primarily interested in *individual* level or *institutional* level measures, we have to begin with the problem of measuring the performance of individuals.

We have spoken of what is to be measured as performance or achievement because this, as we shall see, is what we actually attempt to measure in practice. But levels of performance are determined by a great variety of factors, some of which may be regarded as more fundamental and important. A parent choosing a school may regard the school's position in the league table as saying something about the quality of its educational provision. A university selecting a student may look at that individual's performance and interpret it as a measure of innate ability or potential. In both cases the data on which the choice is made is the same, yet they are being taken as indicative of quite different things. In reality of course, we know that things are much more complicated. Not only are there other factors which affect performance, such as home background, but those factors may interact with one another, meaning that the effect of one will depend on the others which are present. Thus whatever more fundamental latent variables underlie the score which traditional testing produces, their interpretation is not likely to be straightforward. There is an excellent account of the way in which factors of this kind interact in Johnson

(1997) and the ensuing discussion. This is set in the context of the grade point average system used in American universities but the issue at stake is of much wider relevance.

With that warning of difficulties which lie ahead, we now turn to consider how we can extract a scale of measurement from the assortment of test scores which form the empirical bedrock of the exercise.

The Common Sense Approach

Suppose that n individuals are required to take an examination consisting of p questions, or items. A mark of some kind will be awarded to each answer and the results may then be set out in a table as follows.

There is one column for each item and one row for each individual. (Henceforth we use the terms *item* and *individual* as generic terms). The x 's in the body of the table represent the marks awarded and their subscripts specify the row and column, respectively, in which they occur. In the simplest case they may record whether the item was right or wrong; 1 if the answer is correct and 0 if it is wrong. Or they may be marks awarded out of 100, or 20. The items may be questions in a degree examination paper, in which case their number may be as small as 4 or 5, or they may be multiple choice questions running to 50 or even a 100. The items may be sub-divided according to subject or date taken. Individuals may be classified according to school, age or any number of other relevant attributes. Some of the cells may be empty, as when candidates are not required to answer all questions. However, the important thing for our purposes is the near universality of this method of setting out the results of examinations. There must be countless

Table 1. The layout of a typical test result.

Individuals	Items				
	1	2	3	p
1	x_{11}	x_{12}	x_{13}	x_{1p}
2	x_{21}	x_{22}	x_{23}	x_{2p}
3	x_{31}	x_{32}	x_{33}	x_{3p}
.
.
.
.
n	x_{n1}	x_{n2}	x_{n3}	x_{np}

examples in the records of teachers and examining boards of all kinds. The information about performance is contained in the numbers in such tables and the statistical problem is how to extract it.

The following questions arise.

- 1 How should the items be selected—how many and of what kind?
- 2 What population of individuals are we interested in? Should we test them all or only a sample? If the latter, how should it be selected?
- 3 How should the numbers in any row of the table be combined to give a valid measure of the individual's performance?

There are other questions, such as what we can learn from the numbers in the columns of the table about the suitability of the items (e.g. were they too hard or too easy), but we shall concentrate on the three listed above and, especially, on the last.

The choice of items is usually a matter for the judgement of examiners who would be expected to ensure that the syllabus was adequately covered, that the items were of the right 'standard', that they covered the range of skill expected, and so on. Selection of individuals would only arise in large scale investigations, perhaps on a national basis. But if there is no sampling in that sense there are more subtle questions concerning the variation in response one might obtain from one occasion to another with the same individual.

So far as combining the scores in any row of the table to give an index of performance is concerned, the usual practice is to add them up. The row totals, or their averages, have traditionally been seen as the main, if not the only, relevant summary measurement. In the 'right'/ 'wrong' case the total is simply the number correct. Sometimes this may be modified by weighting the individual marks before totalling them or, for example, by selecting only the best q out of p . But that adding up is the sensible thing to do seems to be a matter of common sense.

The central question to be considered later is whether this intuition is well-founded but, in the meantime, a bridge to the theory to follow can be established by probing the matter a little further. We recognise that testing is an uncertain business. We can only use a relatively small number of items in the test and these will, inevitably, give an incomplete picture of the individual's knowledge. Individuals have good and bad days and their answers will reflect not only their knowledge but also their particular circumstances at the time. Response will also depend on the quality of teaching, general educational provision and examination technique. Any particular x in the table can thus be thought of as partly

measuring what it is intended to measure and partly the effects of all the extraneous and irrelevant factors such as those listed above. If we think of these two components as additive we are saying that

$$\text{observed score} = \text{true score} + \text{deviation.}$$

Our intuition that it is a good thing to add up scores then derives from the fact that we feel the ‘deviations’ ought, in some sense, to ‘cancel out’. The more items we have, the more confident we are likely to feel that the vagaries of the examining procedure will be damped out. Crude though it may seem this simple idea, that an observed score is made up of two parts, a real ‘signal’ and irrelevant ‘noise’, is the kernel of the modern statistical approach to the measurement of latent variables.

The Modelling Approach

We now explain how a probability modelling approach tackles the problem. Our aim will be to give a non-technical account which conveys the basic ideas without resort to mathematics. Those who can cope with the mathematics will find a more adequate treatment in Bartholomew (1996) or Bartholomew and Knott (1999).

A statistical model is a mathematical specification of the way in which the observed data are supposed to have been generated. In the present case it will have to describe the linkages between the observed scores—the x 's—and the underlying latent variable. It will need to be a probability model in order to incorporate the element of chance which interposes itself between the true performance level of the individual and what we observe. Once this is done it is a matter of statistical routine to infer what can be said about the latent variable.

It is important to understand what the role of the mathematics is. It does not add anything to the data. Rather it expresses in precise terms how we conceive the data to have arisen and places at our disposal a powerful tool for consistent reasoning.

The logic of a modelling exercise is as follows.

1 We entertain the possibility that individuals may be meaningfully located at points along a scale which we identify with level of performance.

2 We specify a plausible mechanism linking an individual's position on the scale with their scores on the items in a test.

3 We deduce observational consequences of our assumptions under 1 and 2 in order to check whether theory and observation agree.

4 If they do we proceed, as if the model were correct, to make deductions about where to place an individual on the scale given their score.

The model tells us what scores to expect for a given value of the latent variable. We then use probability theory to reverse this process and deduce what latent scale position we would expect for a given set of scores.

There is a weak link in the logic of using a model in this way which is often overlooked but which is particularly important in latent variable modelling. Even if the predictions of the model agree well with observation it does not follow that the model is 'true' in any sense. All that we can say is that people behave *as if* it were true. It may be that there are other models which make the same, or almost the same, predictions. If that is the case we have no empirical way of distinguishing between the competing models.

There are many models used in the field of measurement and their forms depend on the kind of data we have. For example, *item response theory* (IRT) models are for binary items where answers are simply right or wrong. Such a model specifies how the chance of getting an item correct depends on the ability of the individual. Again, if the test scores are expressed on a continuous scale, it may be appropriate to treat them as normally distributed, and then to assume that their relationship with ability is that of simple linear regression. In that case we have a special case of *factor analysis*. The only evidence we can have for any such assumption is obtained retrospectively when we check whether its consequences are borne out in practice.

Given the uncertainties, we cannot expect to determine an individual's position on the latent scale precisely. Within the modelling framework we express our uncertainty about the scale value by a probability distribution and, in particular, by its location and dispersion. Thus we aim to say that if a person's test scores are such and such then we 'estimate' their true position to be P , say, with standard deviation S .

It turns out that we can only do this if we know the distribution of the latent variable in the population from which our n individuals are drawn (known as the *prior* distribution). In the nature of the case this is impossible because, if we cannot even observe the latent variable, we certainly cannot know what its distribution is. We therefore appear to

be at an impasse but, rather surprisingly, it can be circumvented. If we make rather weak assumptions about the class of models to be used—which includes most of those in common use—nearly all of what we need is unaffected by the choice of prior distribution. In fact, we can invoke the statistical notion of *sufficiency* to show that the prior distribution needed for determining the scale value (called the posterior distribution) depends on the observed scores only through a single statistic. In other words, all the information in the observed scores about the latent variable is contained in this single function. It is remarkable that for one of the main IRT models, as well as for the factor model and many others, this single function turns out to be a weighted sum. There is, therefore, a theoretical basis for what common sense seemed to require. But beyond that the theory also provides a means of determining what the weights should be.

Another consequence of the general theory is that, because of the arbitrariness of the prior distribution, we can only provide empirical justification for the rank order of individuals on the latent scale—not the distances between them. This is the price we have to pay for not having to specify the prior distribution. But a little reflection will show that this should not surprise us. If our result does not depend on the choice of prior distribution, then it must be true whatever prior we happen to choose. One prior can be transformed into another by stretching or shrinking the latent scale. Such a transformation will leave the rank order unchanged but not the spacing between individuals. Rank order is thus invariant under changes in the prior. In practice it is usual to assume that the prior distribution is normal. There is nothing to prevent us adopting this or whatever other distribution we please, and the spacing that goes with it, so long as we remember that this is a matter of *convention* and not something which has, or can have, empirical support.

All of the foregoing pre-supposes that a model depending on a single latent variable will fit the data. For this to be true, the items must have been expertly constructed to depend on only a one-dimensional scale of ability. For relatively simple skills this may be possible and there are, in fact, many situations where it has been achieved. But with more general abilities, of which 'general intelligence' is the prime, but by no means the only example, this is usually not the case. In such cases the failure to obtain a good fit can best be explained by the presence of other latent variables. Once more than one latent variable is admitted, a new source of arbitrariness arises, which is at the heart of much of the controversy

about the validity of the whole approach. We shall illustrate the position by means of an example in Section 6 but it may help at this stage to use an analogy to give some idea of the point at issue.

Position on a map is a two-dimensional thing. We can describe the position of Birmingham as so many miles north of London and so many miles west, using the familiar rectangular co-ordinates. But the choice of north-south and east-west as lines of reference is quite arbitrary. We could have used NE-SW and NW-SE axes. Or we could have specified Birmingham's position in terms of a direction and the distance as the crow flies. All such methods enable us to fix its position but none has any claim on us beyond practical convenience. The existence and position of Birmingham is real enough, but the axes we construct to specify position are merely that—constructs.

If our model fitting exercise tells us that we need two dimensions to describe the latent variation of individuals, we have a similar situation. The variation is real, but to define an individual's location we must construct arbitrary axes with reference to which that location is specified. In a geographical context the north/south axis may have physical significance as, for example, if we wished to specify the positions of the towns on the main east coast railway line between London and Scotland. Similarly a particular axis may have meaning in other contexts but that must not be confused with empirical support.

Once we add a second or further dimension, we may no longer be able to rank individuals. The situation is no different from when we are dealing with observable variables. Suppose candidates A and B take two examination papers, and that A's marks are 87 and 64 and that B's are 76 and 56, then we can say that A is better than B. But if B's marks had been 76 and 65, no such ranking is possible. On Paper I alone A ranks higher, but on Paper II the reverse is true. To arrive at a ranking we have to assign relative weights, explicitly or implicitly, to the two papers, and there is nothing in the data to tell us what those weights should be. Equal weights would put A ahead, but by giving Paper II sufficient weight, B could be made to come out first. A one-dimensional ranking of locations in a many-dimensional space is not, in general, possible. All rankings, such as degree classifications, therefore depend not only on the marks obtained in the examination, but on independent judgements of the weights to be attached to the papers. The same is true of latent variables. We shall illustrate these points by examples in the following sections.

Statistical versus Psychometric Inference

In the psychometrical literature a distinction is drawn between these two kinds of inference. The point at issue can be explained by reference to the array in Table 1. In psychometrical inference the interest is in generalising from the results for the p items selected for the test, to the universe of items which might have been used. The items used are viewed as a sample of the domain of knowledge being examined, and thus subject to uncertainty. The presumption is that we could get nearer to the 'true' ability of each individual if we could increase the number of items indefinitely.

In statistical inference, on the other hand, the n individuals are regarded as a sample from a larger population to which we wish our inferences to apply. The larger the sample size, the more precise will be our information about the characteristics of the items which happen to be included in the test.

In practice both types of generalisation will be of interest, though one or other may be uppermost. For inter-school or regional comparisons, for example, using the same set of items throughout eliminates variation arising from item selection and thus makes school comparisons more efficient. For establishing rank orders within a particular group of individuals, however, what matters is the coverage of the field of knowledge.

Attempts at finding a theoretical basis for psychometrical inference have not been entirely convincing. The root of the difficulty lies in the fact that the domain of items is rarely well-defined and, even if it is, the selection of items is not usually made in a well enough defined way to allow valid generalisations. All that we can do then is to look at the variation among the items that we have selected. For example, if in Table 1 the column totals were identical, we might conclude that all items were equally difficult and hence that the results could be safely generalised. If, on the other hand, there was a good deal of variation in the column totals, we would be less confident that a similar ranking would be obtained among individuals if different items were to be used. All of this pre-supposes that the actual items used are, in some sense, representative. In most cases it is the lack of any information about the sampling process for items which precludes valid inference beyond the set of items to hand.

Although this traditional distinction may be helpful in clarifying our thinking, it is unnecessary in practice. Our modelling approach enables

us to make general statements about the latent variables without the need to specify how the items were selected. All that we need to do is specify how those particular scores were generated by the latent variable. This tells us what can be validly said about the latent variable *given* the items which were selected. Equally, if the n items are randomly selected from some larger population (real or hypothetical), we can use traditional statistical inference procedures to generalise to that population.

An Example Illustrating Indeterminacies in the Model

So far we have described what a model aims to do and have made some assertions about the practical consequences. These can only be fully justified by mathematical analysis but they can be illustrated by examples to make their implications clearer.

The first example has a long history and was deliberately chosen for that reason. It consists of a test of five items, scored right or wrong, administered to 1000 individuals. It is Section VI of the Law School Admission Test. The left-hand column of Table 2. lists all the 32 possible outcomes of the test ranging from 00000 for someone who gets all items wrong to 11111 for someone who gets them all right. The second column gives the frequencies with which each response pattern occurred. It is immediately obvious that some response patterns are very much more common than others; 173 individuals produced 11011 and only one 00100, for example. Our aim is to try to understand the reasons for this variation and thus to discover what can be inferred about the abilities of the individuals.

We may start with the hypothesis that there is no underlying variation in ability whatsoever. After all it is pointless to enter into debate about scaling ability if the test is no better than a lottery. If this were really the case, all individuals would have the same chance of getting any particular item correct and, let us assume, the chance of getting any item is independent of whether other items were right or wrong. If all of this were true, we can work out what the frequency distribution of response patterns would be. The result is given in column 3 of Table 2. Some are very close to the observed frequencies, but there are also some big discrepancies, most notably where 4 or 5 items are answered correctly. A formal test of agreement confirms that the observed frequency distribution is most unlikely to have arisen from the simple model just described.

Table 2. Observed and expected frequencies for various hypotheses about the distribution of the latent variables for the Law School Admission test Section VI.

Score pattern	Observed frequency	Expected frequencies when:		
		No variation in ability	Two-point distribution	Normal distribution
00000	3	0.3	1.6	2.3
00001	6	2.0	5.6	5.9
00010	2	1.0	2.5	2.6
00011	11	6.6	9.3	8.9
00100	1	0.4	0.7	0.7
00101	1	2.5	2.7	2.6
00110	3	1.2	1.2	1.2
01100	0	0.9	0.9	0.9
00111	4	8.1	5.9	6.0
01000	1	0.7	1.8	1.8
01001	8	5.0	6.7	6.4
01010	0	2.4	3.0	2.9
01011	16	16.0	13.6	13.6
01101	3	6.1	4.3	4.4
01110	2	3.0	2.0	2.0
01111	15	19.8	14.0	13.9
10000	10	3.7	9.4	9.5
10001	29	24.8	35.4	34.6
10010	14	11.9	16.1	15.6
10011	81	79.8	75.5	76.6
10100	3	4.6	4.7	4.7
10101	28	30.7	24.5	25.0
10110	15	14.7	11.2	11.5
10111	80	98.7	84.6	83.5
11000	16	9.0	11.6	11.3
11001	56	60.3	55.3	56.1
11010	21	29.0	25.3	25.7
11011	173	194.4	174.2	173.3
11100	11	11.1	8.3	8.4
11101	61	74.7	63.6	62.5
11110	28	35.9	29.6	29.1
11111	298	240.0	294.8	296.7
Total	1000	999.3	999.9	1000.2

The second model, which is one of those widely used in item response testing, supposes that individuals vary in ability scaled in such a way as to make its distribution normal. In this case the predicted frequencies are as given in the last column of Table 2. The agreement is very much closer, and it may be shown to be well within the limits one would expect if the model were true. This, quite reasonably, has been

held to justify regarding the test as measuring variation in the ability needed to pursue a course in law. But that is not the end of the story as the penultimate column in Table 2. shows. This is calculated on the hypothesis that the individuals are divided into two latent groups—what we might call ‘high’ and ‘low’ ability groups. A close comparison of the two columns shows that there is hardly anything to choose between them. We can certainly conclude that there is evidence of variation in ability because both models which incorporate this feature do very much better than the one which does not. But when it comes to distinguishing between the radically different patterns of variation in ability which underlie the alternative models, the data give us no practical help!

If we choose to go ahead with the model of continuous normal variation, on the grounds that it is fully consistent with the data, there are further problems. We pointed out that the model enables us to predict an individual’s location on the latent scale and to specify our uncertainty about it. The latter can be done in terms of the variance. Prior to observing the test results, the uncertainty may be expressed by the variance of the prior distribution. After we have observed the responses it will be the variance of the posterior distribution that is relevant. In the case of this example, the latter figure is about two thirds of the former. In other words, this particular test has not greatly reduced our uncertainty about where the individual lies. The way to obtain more precise information is to add more items, and theory could guide us on how large the test should be. But the example warns us that a valid test may not be a reliable one.

An Example Illustrating a Two-dimensional Latent Variable

This example relates not to test scores but to the related question of the anxiety people feel about taking tests. It has been chosen because it has been used in many countries (Norway, Germany, Holland, Egypt, India, Hungary, Spain, Korea and Canada), with male and female respondents and with very similar outcomes in all cases. It therefore appears to describe something which is more than an artefact of a particular set of circumstances. We use Canadian data for which further details will be found in Gierl and Rogers (1996). The Test Anxiety Inventory consists of 20 questions about how people feel before taking an examination. They are listed in the Appendix. Individuals respond on a 4-point scale which, for the purposes of this example, are treated as

points on a continuous scale. They are usually analysed using a factor model which supposes that the response is a linear combination of one or more latent variables together with an 'error' term. It appears that two latent variables are needed to explain the response patterns and, therefore, that 'test anxiety' is not a one-dimensional phenomenon. This raises the question of how we interpret these dimensions and what reality they have. Some results are given in Table 3.

Readers familiar with factor analysis will recognise that the numbers in this table are *factor loadings*. For our limited purpose here it is sufficient to know that they may be interpreted as correlation coefficients between the item scores and the latent variable supposedly underlying them. Thus in column I of the orthogonal set we notice that all of the correlations with the first latent dimension are positive, mainly large and, for the most part, roughly equal. Since all of the items express anxiety in some form it is natural to identify the dimension with the 'test anxiety' which the items are supposed to be measuring. But that is not the whole story, because there is a second factor represented by

Table 3. Alternative sets of factor loadings for the Test Anxiety Inventory for 389 Canadian female students.

Item	Orthogonal set		Oblique set		Emotional (E) Worry (W)
	I	II	I	II	
1	.64	.04	.41	.29	—
10	.55	-.04	.44	.15	E
11	.80	-.08	.66	.18	E
12	.73	-.05	.58	.20	—
13	.61	-.07	.51	.13	—
14	.61	.31	.05	.65	W
15	.81	-.19	.74	.05	E
16	.68	-.24	.77	-.07	E
17	.65	.28	.11	.62	W
18	.62	-.09	.54	.11	E
19	.58	-.37	.45	.16	—
2	.72	.07	.43	.36	E
20	.66	.22	.20	.54	W
3	.46	.26	.01	.52	W
4	.62	.18	.22	.47	W
5	.49	.35	-.09	.66	W
6	.44	.28	-.03	.54	W
7	.63	.23	.16	.55	W
8	.71	-.22	.77	-.04	E
9	.69	-.33	.89	-.20	E

column II. This is independent of the first—hence the term orthogonal—and much more weakly related to the observed item scores. The interesting thing is that some of the correlations are positive and some negative. Some help in interpreting this dimension is provided by the last column in which most items are classified as ‘worry’ and ‘emotional’ items. The latter are distinguished by the occurrence of autonomous nervous system reactions as, for example in the statement ‘While taking examinations I have an uneasy, upset feeling’ (item 1). An example of a ‘worry’ item is ‘During exams I find myself thinking about whether I’ll ever get through school’ (item 5). Four items were not readily classified in either category. For the most part emotional items are negatively correlated with the second latent variable, and the worry items are positively correlated.

The conclusion which emerges from all of this may be expressed as follows. The pattern of responses can be explained by supposing that individuals vary in two dimensions. The dominant dimension can be identified with a generalised kind of anxiety in examination situations of the kind indicated by what is common to the 20 items. But, given any position on this axis, there will be a more limited variation in a direction independent of the first, which distinguishes those where the preponderant aspect is emotional from those for whom it is cognitive.

However, analysts have usually chosen to identify the ‘worry’ and ‘emotional’ aspects as the more fundamental. That is they have chosen axes to which individuals are referred which corresponded with these two supposed variables. The results are shown in columns I and II of the ‘Oblique set’. According to this representation factor I is the one which, predominantly, correlates with the emotional items (picked out in bold type); factor II correlates with those in the worry category. We then conclude that individuals are characterised by where they stand on those two scales. Unlike those in the Orthogonal set, these latter scales are not independent, but are typically correlated to the degree of about 0.7.

We now have two descriptions (among many others that are possible) of the dimensions of latent variation. One can be generated from the other by a process known technically as rotation. Since there is no empirical means of choosing between these two descriptions it is often argued that neither is real and that it is futile, if not actually harmful, to speak as if there were. This objection misses the point. Both representations describe the same thing but in different ways. Each may be useful in some circumstances but not in others. For example, if our purpose is to understand what makes people anxious about exams, it may be

useful to distinguish those feelings which have physical correlates in the functioning of the nervous system from those which do not. If we are more interested in distinguishing those individuals for whom pre-exam anxiety might be an important determinant of performance, it is more useful to know that most of the relevant information is conveyed by dimension I of the orthogonal set. The reality is that individuals differ. The arbitrariness lies in what axes of reference we use to describe those differences.

Group Comparisons

With the advent of league tables of various kinds, comparisons between groups and over time have been at the centre of arguments about the measurement of standards. Because of the public debate on the matter, the problems of interpretation are widely known. The annual dispute about whether improved A-level or GCSE results are due to falling standards or to harder work, better teaching and so forth, serves as a reminder of the issues involved. Here we review these controversies in the light of the modelling approach advocated in this paper.

The essentials can best be exposed by using a little elementary algebra. Suppose that an individual's examination mark on a particular item is denoted by X . We suppose that this depends on how easy the item is, how good the item is for discriminating between people of differing ability, how able the individual is (or how well taught, favourably supported by home environment etc.) and on a multiplicity of minor factors peculiar to the item, the individual and the circumstances. A simple representation of X is then

$$X = E + DA + M.$$

E measures the easiness, and the bigger it is the larger will be X . A is the compound measure we called ability. How much effect this has on X depends on the size of D , the discrimination factor. If the item is good at discriminating, a small change in A will produce a large change in X . Finally, M represents the combined effect of all other factors.

The nub of the difficulty in making comparisons arises from the confounding of D and A and there is no way we can separate the two effects. The usual way out of the impasse is to standardise the distribution of A . We have already noted that this distribution is arbitrary and, as a matter of convention, we can give it any origin and scale that we please. The usual convention is to scale the distribution of A to have zero

mean and unit standard deviation. But the very essence of the problem of making comparisons is that the distribution is not the same for the two populations. If we fit the model to data from two groups, any difference in the abilities will show up in the parameters E and D which characterise the items. But if the items are the same, then the difference must be attributed to A . Thus valid comparisons can only be made if the institutions are the same in all other respects. In particular the factors represented by M must be the same. If environmental and other background factors are not the same, this assumption will not be valid. The problem is particularly acute when making comparisons for the same population over time. Performance, or standard, is not the only thing which varies over time, and there is no way the effects can be separated.

All of this is predicated on the supposition that the equation adequately captures the way in which the various factors combine to produce a score. It does not, for example, allow for any interaction between items and institutions. A blatant case of this would occur if one school 'taught for the test' and the other did not, but the effects might be more subtle. All group comparisons must be qualified by the statement 'other things being equal' and in the nature of the case that is a judgement which cannot be tested empirically.

Conclusions and Criticisms

It will be clear that the modelling approach highlights the hazards of attempting to measure standards. Even if a single latent variable is adequate, the most that we can do is to justify a ranking of individuals and, unless the number of items is large, that is likely to be subject to a high degree of uncertainty. When additional latent variables are needed, there is a fundamental arbitrariness in describing the latent space which complicates interpretation. Comparisons between groups or institutions are fraught with difficulties caused by the fact that we cannot separate out all of the factors which contribute to the scores of individuals. This might seem to argue for the abandonment of the enterprise altogether, or reversion to the simple practices which have served for generations. This is to mis-read the situation. The cruder methods do not avoid the difficulties, they merely ignore them. The virtue of the approach we have outlined is that it makes explicit what would otherwise be only implicit. Our intention is constructive not destructive; by showing what is indefensible, the way is cleared to build on a more secure foundation.

There have, however, been criticisms of the whole approach which, if accepted, would leave little intact. These have centred largely on the concept of general intelligence but they apply here also, if with less force. Two eloquent critics are to be found in Gould (1984) and Rose (1997). Neither author is a social scientist nor an expert in measurement theory and their accounts are not wholly reliable. Nevertheless, as popular writers whose work reaches a wide lay audience, it is their interpretation which most people are likely to have encountered.

Each critic seeks to demolish the notion of any real thing called general intelligence using facts which have already come to light in the course of this paper. Rose correctly draws attention to the arbitrariness of the form of the distribution of the latent variable, though it is unclear whether he is actually referring to the latent variable itself, or to some score (such as the sum) derived from the item scores. There is certainly no requirement that this should be normal and the only inferences which are legitimate are those which are independent of the form of the distribution. The main attack of both authors is reserved for the arbitrariness of the axes in the factor space. If different factors can be made to come and go by the whim of the investigator in rotating axes, they cannot be claimed to be 'real'. Spearman's general intelligence vanished when Thurstone rotated the solution to produce a cluster of specific factors. In our example, test anxiety dissolved into two correlated dimensions labelled 'worry' and 'emotional'. The axes we used are, indeed, arbitrary but the space which they span is real in the sense that it is a collective property of the set of items. Whenever several dimensions are needed, no absolute ranking of individuals is possible. It is therefore perfectly legitimate for Rose to point out that different rankings will result from different tests. What matters, however, is relevance for purpose, and we indicated in the test anxiety example how different axes might usefully serve different purposes. A full critique of Rose and Gould is beyond the scope of this paper. Our purpose in raising the matter here is to make clear that nothing they say affects the central argument of this paper. It is variation between individuals which is real; how we describe it is arbitrary but not meaningless.

Appendix

Topics of the 20 test items for the example of Section 6:

1 Lack of confidence during tests

- 2 Uneasy, upset feeling
- 3 Thinking about grades
- 4 Freeze up
- 5 Thinking about getting through school
- 6 Harder I work, more confused
- 7 Thoughts interfere with concentration
- 8 Jittery when taking test
- 9 Even when prepared, get nervous
- 10 Uneasy before getting the test back
- 11 Tense during test
- 12 Exams bother me
- 13 Tense/stomach upset
- 14 Defeat myself during tests
- 15 Panicky during tests
- 16 Worry before important tests
- 17 Think about failing
- 18 Heart beating fast during tests
- 19 Can't stop worrying
- 20 Nervous during tests, forget facts

Discussion

Harvey Goldstein

Introduction

The term 'standard' has come to mean in many educational systems a position on a measurement scale, either ordinal or continuous and presupposes the ability to construct measurements, especially those based around student achievements, along such a scale. Unfortunately, discussions about 'standards' of educational attainment have suffered from the absence of a widely acceptable formal framework. Yet such a framework, with clear definitions and rules for deriving conclusions from stated assumptions, is both desirable and necessary for informed debate. In this paper I shall explore the concept of 'educational standards' and examine the formal assumptions which underlie it.

It appears that the only sustained attempts to provide a formal framework have been those of psychometrics. Yet this discipline has

been concerned principally with providing mathematical *models* to describe the responses of subjects (people) to test questions or items, and in particular ways of achieving efficient summaries of those responses. It seems to have been content with establishing a mathematically consistent structure for this special case rather than attempting to formulate a general structure which would allow a wider debate to emerge. For a discussion of some of the limitations of the psychometric approach see Goldstein and Wood (1989). David Bartholomew has provided a succinct discussion of educational measurement and has emphasised the importance and to some extent, arbitrariness, of the assumptions that have to be made in terms of choice of items and populations.

I wish to explore different kinds of frameworks which may be used to characterise notions of educational standards. I start with a definition of a simple formal structure, then describe an alternative approach and finally look at some practical implications. My principal argument is that, without a formal framework, it is extremely difficult to have a useful debate about *measuring* educational standards. The following suggestions are an attempt to provide such a framework in the hope that this will stimulate further debate.

A Simple (Constructionist) Definition

Consider a simple case: we are measuring the attainment of a population for a class of arithmetical operations. Assume that the population is well defined (e.g. all year three children in Welsh schools) and that the class of arithmetical operations is also well defined (e.g. the addition of two 2-digit numbers). Assume also that a suitable sampling procedure is available for the population which will enable an estimate to be made, say of a population mean, with a predetermined accuracy.

In order to provide measurements a procedure is required for constructing a measuring instrument, a test. To construct this test we need to select the items (in the case of the above example there is a finite population of such items), determine how they are presented, administered and the children's responses assessed. Research (see for example Foxman *et al.* 1990) indicates that variations in item format and presentation can affect the proportion of correct responses, so that I shall also assume that these aspects are systematically controlled.¹

¹ Note that other factors such as the ordering of items may affect responses.

Note that other sources of (random) error can arise from the requirements of particular procedures, for example if items are selected from a population of items. Any sampling error produced in such ways can then be incorporated into statements about the precision of estimates.

I shall also suppose that the test constructors have been very careful to carry out a detailed conceptual analysis of what they are trying to measure and that the test and its items have been carefully piloted.

The simple definition is just the population mean of a measuring instrument defined as above. Changes in the standard are measured by the difference in means across populations or across time. A formal definition for this simple case is given in the appendix.

Implications of the Simple Definition

We note that this simple definition implies some important restrictions. First, it requires a very precise operationalisation. Inferences are applicable only to the class of items we have defined. Thus, keeping with our example, if we wish to study 2-digit arithmetic addition when the items are presented in a different fashion, say 'horizontally' rather than 'vertically', then we will need to construct a different test which will then refer to a different 'standard'. This immediately raises the issue of the relationship among such standards. The study of such relationships is then a matter for empirical research and will not be dealt with further in the present paper.

Secondly, it assumes that there is a very clear definition of the population or 'universe' from which the items are selected. In our simple example such a definition seems possible, but in more complex cases, for example in a test of reading comprehension, it may be impossible to define the universe precisely other than in the degenerate case where it consists simply of the single test being used. It is also possible to construct a set of tests and then to make random selections from this set. Such a procedure is similar to those used in conjunction with test equating (see below).

Thirdly, our definition is purely formal. It says nothing about relevance or appropriateness of any test we may construct, nor does it say anything about how a universe of items or a relevant population is to be chosen. As we will argue below, it is just when such choices have to be made that important problems arise which generally cannot be solved by a purely formal procedure.

Applications of the Simple Definition

The simple definition has an obvious application when the same test is used with different populations, usually defined with respect to time (see Start and Wells 1972 for examples). A variant is where a common set of items is used within two or more different tests and the common set forms the basis for inference about population differences.²

Problems arise when we wish to make interpretations of any differences: the same kinds of problems arise if we sample from a well defined universe of tests. Start and Wells (1972) show how a reading test used in the late 1940s had changed its interpretation by the 1960s. Language and curriculum usage had changed over the period and the test had become 'harder' as component items became less familiar to the students responding to the test. We may refer to such changing conditions as 'background conditions'. Changes in mean response were observed but it was not felt possible to separate a changing 'test difficulty' from a 'real' population response change.

Over a short period of time, if it is assumed that changes in background conditions are at most negligible, we may draw inferences about standards in terms of the above definition, using either a constant test or one sampled from a suitable universe. Note, however, that we need to make further assumptions, about background conditions, in order to reach conclusions about 'standards'. In the following sections we investigate ways in which changes in such assumptions will affect the inferences which can be made. To introduce this discussion we introduce an underlying philosophical distinction.

Platonic Standards

Most tests, and for that matter examinations, are not constructed using a systematic sampling procedure. Tests are usually constructed by carefully following criteria to do with content, format, relationships with other measurements, tests or items, group differences (biases) and so forth. Empirical piloting and expert assessment may also be applied before a test is used. This procedure may be referred to as 'Platonic' test construction since it relies upon the notion that there exists an 'ideal' underlying test item universe or concept and that the procedures used to

² For a detailed discussion of such a procedure see Beaton and Zwick (1990). It raises practical difficulties of interpretation, however, since the 'context' of the items differs according to the test within which they are used.

construct a real test are realisations of it. The construction 'rules' are designed to sample from this ideal: unless the concept of such an ideal is invoked the construction rules will be arbitrary. This underlying concept typically will be referred to as 'reading comprehension', 'arithmetic ability', 'understanding of mathematical symmetry' and so on and it is clear that the aim is to make statements about the underlying concept. This Platonic procedure of assuming an underlying concept is different from the constructionist procedure outlined in the first section which formally defines either a particular test or an explicitly constructed universe as the object of inference. Of course, a constructionist procedure may draw upon Platonic notions in order to derive a population in the first place, but thereafter it relies upon clear sampling rules for its operation.

The distinction we are making is more than a philosophical nicety: it has profound implications for the kinds of interpretations that can be made. Consider the above example of the reading test which was assumed to become outdated over time. Once the test had been devised, a strict constructionist interpretation would be indifferent to issues such as 'relevance to curriculum'; all that would be required is that the rules for administering the test to successive populations be adhered to. The additional observation that the test was less relevant in the 1960s can only be admitted if the test itself is just one instance of an underlying reality, which means the adoption of a Platonic viewpoint. Even in the case where the test was used over a short period so that 'relevance' could be assumed to apply, we would still be appealing to this additional assumption of relevance to justify the use of the test.

It would seem that the Platonic view of a test is the one that is very widely adopted, although not universally. The constructionist example we started with, of a well defined arithmetic test, might be regarded as useful if our interest centres on just the universe of items sampled by the test.

One possible objection to the distinction we are making is that the constructionist procedure can be extended to incorporate many of the tests people use without necessarily invoking a Platonic viewpoint. Thus, for example in the reading test case we can envisage a 'super-universe' of items which is constructed by considering the union of the sets of items relevant to all possible populations. If we suppose that it is possible precisely to define a universe for any given population, say of arithmetic or spelling items, then we need merely form the union of these separate universes and sample from this. This presents several

difficulties however. First, it does not deal with the example we used to introduce this section, which is very commonly used. Secondly it requires an assumption about the weight to be attached to each item in the super-universe for the purpose of sampling. In general the different populations will have items in common: if we have two populations are the common items to be counted once or twice in the super-universe? Thirdly, if the populations are those chosen at different times, the sampling at the first time can only be made from the universe of items defined at that time—subsequent times are not yet observed and their universes cannot be defined. This introduces an asymmetry which prevents us sampling from the same super-universe at each occasion.

We shall not consider in any more detail the constructionist procedure: the social and political debate on standards rarely is concerned with these. In the following sections we will pursue some of the implications of the Platonic approach.

Platonism in Practice

To understand the implications of a Platonic view of standards, we shall look in a little more detail at that part of mathematics sometimes referred to as 'numeracy' which is largely a subset of elementary arithmetic. Specifically we shall contextualise what we have to say within the political requirement set out by the British Government regarding 'standards of achievement' for Key Stage 2' (eleven year olds) children over the period 1997–2002 (DfEE, 1997). This example is chosen because it illustrates the issues in a straightforward fashion and because it has some important contemporary educational implications.

Broadly speaking, attention is focussed on the percentage of children in England and Wales achieving a 'level' 4 in numeracy tests. From a national mean of about 55% in 1997 it is proposed that this should rise to about 80% by the year 2002; several educational programmes have been devised in an attempt to achieve this target. The level for each child, on a scale from 1 to 10, is assigned on the basis of responses to a test, separate tests being devised each year.

It would clearly be possible to adopt a constructionist procedure, whereby an item universe was defined at the outset and sampled from each year. This, however, would appear not to be under consideration and would be difficult to adopt since over several years the definition of the relevant universe would almost certainly change. Rather, the term numeracy is being used in what we have termed a Platonic sense. It is assumed to

exist, independently of any specific version of a curriculum, and there are assumed to be more or less well defined procedures for constructing tests which reflect it. Assuming that it does exist, how are we to ensure that our succession of tests reflects it (and no other concept)?

Suppose that we are confronted with different tests at each occasion and that we are not relying upon a common set of items to provide 'continuity', for the reasons outlined above. The possibility of more general 'equating' procedures³ arises but is not relevant, for the following reason.

The aim of test equating is to allow the scores on two tests, say test A and test B, to be calibrated along the same scale. This may be attempted in a number of ways, but for our purpose suppose that it is done by independently giving each test to each member of a suitably large random sample from a population. Suppose also that the equating 'works' in the sense that each test ranks the sample in a common ordering so that a unique correspondence between test scores can be set up. If we apply such a procedure to two tests at different time points we immediately face the problem of which population is to be used to carry out the equating. If it is the first one, then we require the second test to be available at the time of the first test: this clearly requires that all tests are devised at the outset, which is equivalent to defining a super-universe in the constructionist sense and is anyway practically infeasible. If the standardising population is that at the second occasion then we have a similar problem. Thus, unless we define, as before, a super-universe at the outset, we cannot sample from the later populations at the first time occasion.

In effect, equating attempts to make the same inferences as would be made were the same test to be given at each occasion. Thus, the use of equated tests raises no new fundamental issues beyond those discussed when the same test is used in a constructionist sense. In practice it also introduces further 'noise' since no equating is perfect and there is the problem that the equating calibration relationship may differ across subpopulations of interest (see Goldstein and Wood 1989 for a further discussion).

Having ruled out both simple constructionist and related equating

³ Some kinds of equating procedures, especially those using item response models, are based around such a common item set. Thus they suffer not only from the 'context effect' but also from the problems associated with equating procedures in general (see Goldstein and Wood 1989 for a further discussion).

approaches, what else may be available that allows us to create a series of tests which reflect the concept of numeracy? One procedure would be to invest the responsibility for conforming to a Platonic standard in the hands of a group of individual 'experts' who would use their judgements when creating (and interpreting) tests. Effectively, this is what is done by the British public examination systems where the experts or 'examiners' use a variety of methods, including the study of statistical performance information, to arrive at 'comparability'. It also underlies the various 'standard setting' procedures which are judgementally based (see for example Morrison 1994). It is clear that those involved believe that they are attempting to achieve a correspondence with an underlying or Platonic standard (Cresswell 1997a or b).

An interesting feature of these procedures is that they require a post hoc component. It is not assumed to be possible to create a test or examination, however carefully constructed, that ensures comparability *without* using the empirical evidence obtained from a set of actual responses.⁴

A Basic Limitation of Platonic Standards

Whatever procedure is used to construct a test and to manipulate subsequent scoring or grading systems (a 'testing system'), there is a fundamental problem with the use of a Platonic standard. Any particular testing system will be an approximation to the standard in question. It will be a matter for debate as to how good such an approximation is, and the effort that goes into the construction and scoring is largely devoted to attempting to improve such an approximation.

Nevertheless, *there is no objective way to determine how close any approximation will be*. In particular there is no way of knowing whether the responses to two different testing systems *differ from the standard by the same amount*.⁵ In other words any observed difference (apart from

⁴ It would be possible to obtain a very approximate correspondence to a Platonic standard by removing the post hoc element. This would allow distinctions to be made between 'extreme' performances on the examination and would also allow very crude comparisons over time. In the latter case, however, it would be unable to detect anything other than gross changes.

⁵ As in the examination example, it *may* be acceptable to use such approximations for the purpose of detecting 'very large' changes in an underlying standard over time. There remains the problem of defining 'very large' and in practice what is usually required is the detection of moderate change over relatively short periods.

sampling errors) between two populations with separate testing systems will reflect both any underlying difference and the different extent of approximation for each system. This results in an unresolvable duality. Thus, for example, the requirement to detect a given amount of underlying change, as in the DfEE (1997) case, is unrealisable.

This *duality principle* was referred to briefly when discussing the application of a reading test to populations widely separated in time. In that case the debate centred upon the uncertainty about whether the difference between the test and the underlying standard, that is the size of the approximation, had remained constant over time or changed in a particular direction. In effect it was argued that, since the correspondence to the school curriculum in particular and to language in general had weakened, so the approximation had become worse. If such an argument is accepted then clearly the test could not be used: there existed no measure of how much the approximation may have changed.⁶ As mentioned earlier, only in the simple case where, for example, populations are separated by only a small time difference, may we reasonably assume that the approximations (using the same test) are similar and so attribute even a moderate change in population responses as a change in the underlying standard.

If one accepts the Platonic viewpoint it follows that to make valid comparisons using different measuring instruments it must be demonstrated that the approximations involved have effectively the same magnitude and sign. In doing this one may appeal to the various procedures used and possibly to independent evaluations of them. Any claim about 'standards' then becomes part of a wider concern about the adequacy of the procedures used, and ultimately, perhaps, such a debate may lead to improvements in those procedures.

Comparisons of Educational Systems

There are now many studies involving comparisons of test responses in different countries. This generally involves the same test being used (with translation where appropriate) at a particular time with different educational systems for purposes of comparison. It is not entirely clear whether a constructionist or Platonic view is held by those involved. It

⁶ In this case, even though a very long time period of over 15 years was involved, the observed differences were not accepted as large enough for a change to be detected given the approximations which seemed to be involved.

might seem that a purely constructionist perspective is held in that basic results are typically presented as comparisons of mean scores. In addition, however, it is recognised that responses are related to curriculum content and such relationships are also presented; from such a perspective one might suppose that a Platonic view of different degrees of approximation are assumed to exist. I shall not pursue this further here, but a discussion of some of the most important studies of this kind can be found in an issue of the journal 'Assessment in Education' (Goldstein, H., 1996a).

Implications

My argument may be summarised thus:

1 It is possible to define a constructionist standard for a single test or one derived according to well specified sampling rules. This, by definition, can be used to compare populations and to form (probabilistic) judgements about differences or changes over time. While this approach may be applicable in some circumstances, it appears to be little used in practice.

2 A Platonic standard may be conceptualised, and a testing system can be designed to approximate to it. Such approximation may be adequate for many assessment purposes. When used to measure population *differences*, and in particular changes over time this approach suffers from a fundamental limitation known as the duality principle. This stems from a lack of objective knowledge about the size of the different approximations involved. This limitation precludes the use, in general, of Platonic standards to compare populations unless we can argue convincingly that the approximations used are equivalent in the different populations.

If these points are accepted, then attempts to construct a standard of whatever kind have to confront the difficulties. There appear to be few uncontested examples where a convincing case in favour of a Platonic system has been made. Certainly this case does not seem to be accepted in what is perhaps the most sophisticated large scale system, that of public examinations in Britain.

I have alluded to the possibility that we should restrict comparisons over time, and perhaps across populations, to the detection of gross or extreme changes. In other words we could regard our tests as screening devices for detecting when major changes might be occurring, rather

than as precise measurements. Another possibility is to forsake attempts to measure absolute differences in this way and restrict attention to what might be called ‘second order’ changes, by which I mean the following.

Over time, using a Platonic definition and accepting possibly different degrees of approximation, we can study the relationship between test scores and other factors for each test. Thus, for example, we can examine a gender difference in an attempt to judge whether such a difference had changed between tests. Naturally, we would need to be able to carry out a common standardisation for each test, perhaps simply requiring them to have identical score distributions—this would automatically prevent absolute comparisons but still allow relative, what I have termed second order, comparisons to be made. Such comparisons will be scale dependent and different standardisations may result in different interpretations, but the possibility for potentially useful statements does seem to exist. One might go further and suggest that these second order comparisons are more practically useful than absolute or first order ones since they are an attempt to move closer to causal explanations. Thus, for example, the finding that the difference between girls and boys in examination performance has changed over time (Elwood and Comber, 1996) has generated a debate about possible causes, as well as research into factors which might explain such a change.

Appendix

Formally, denote a specific test for a population as

$$X^{(t)} = \{x_1^{(t)}, x_2^{(t)}, \dots, x_{n(t)}^{(t)}\}$$

where t denotes the target population, $n_{(t)}$ is the number of items in the test and $x_j^{(t)}$ is the j -th item. Denote $\bar{x}^{(t)}$ by the sample estimate of $\mu^{(t)}$, the required population mean.

The simple definition of a difference in standard between population 1 and population 2 is $\mu^{(1)} - \mu^{(2)}$. These populations may be two ‘real’ populations (for example Wales and Scotland) or the same geographically defined population at different times (for example Wales in 1970 and 1990). Typically we shall be considering the latter case. Using sample estimates we wish to make a statement about $\mu^{(1)} - \mu^{(2)}$, for example to provide a confidence interval.

We can extend our reasoning to any population parameter, for

example the median, without alteration. We can also extend our reasoning to *any* well defined procedure for eliciting responses, for example based upon detailed observations or the administration of practical tasks.

Ian Plewis

The 1988 Education Reform Act led to a system of national assessment, with compulsory assessment for pupils in state-funded schools at the ages of seven, eleven and fourteen. More recently, so-called baseline assessment has been introduced for pupils in their first term of the reception year although the instruments used are not uniform across Local Education Authorities. The nature and purposes of national assessment have changed over the last decade but, throughout, there has been less interest in accurately ranking individual pupils than is the case for public examinations at ages sixteen and eighteen. Important as the outcomes of the national assessments are for pupils and their parents, they do not affect pupils' life chances in the same way as GCSEs and A-levels do. The agenda is, therefore, increasingly driven by what Bartholomew refers to in section seven of his paper as 'aggregate' comparisons. And, as his and earlier papers make clear, it is very difficult, if not impossible, to make the comparisons which politicians are demanding from the system of national assessment which they created.

There are, in principle, at least five kinds of comparisons for which a system of national assessment might be used:

- a comparisons between schools;
- b comparisons over time or between cohorts of pupils;
- c comparisons between different subject areas;
- d comparisons of the performance of different socio-economic and demographic groups;
- e comparisons over age or developmental changes.

The first two of these are the ones favoured by politicians. Comparisons between schools in the form of rankings, or league tables, are, despite their popularity with some politicians, now widely recognised to be fatally flawed (see, for example, Goldstein 1999). Analysing differences between schools can, however, form the basis of a useful research agenda.

As Bartholomew points out, comparisons over time rest on dubious assumptions. These seem likely to become increasingly untenable as the stakes attached to results at Key Stages One and Two become higher so that teachers teach to the test more and more. There are also, as Plewis (1999) points out, problems when it comes to assessing performance in different subject areas.

Comparisons of the performance of pupils in different socio-economic groups have received rather little attention, with the possible exception of gender differences. However, if we are prepared to make some assumptions, and ultimately all comparisons rest on assumptions which are often difficult to test, then some progress can be made. In terms of the equation on page 135:

$$X = E + DA + M$$

then, if we assume that *E* (easiness), *D* (discrimination) and *M* (other factors) do not vary across groups, then we can, in principle, look at inequalities. Moreover, if we assume that changes in *E*, *D* and *M* are uniform across groups then we can look at how inequalities are changing. It is important to remember, as Plewis (1998) points out, that overall improvements over time in the proportions of pupils reaching, say, level four at Key Stage Two can be consistent with increasing inequalities.

In many ways, changes with age—or developmental changes—are the changes most closely related to learning and might, therefore, feature more strongly in debates than they do at present. The methodological challenges of constructing a scale applicable over the ages five to sixteen, so that changes with age can be measured, are considerable. On the other hand, this is a potential strength of the ten point scale currently in use. Unfortunately, the absence of any concerted research on the properties of the ten point scale is regrettable. To return to my earlier point about the differences between the purposes and organisation of assessment systems, the public examination system is run on essentially market principles and yet, despite the wish to protect commercial secrets, there is, ironically, more methodological research published from the exam boards than there is from the quango which is the Qualifications and Curriculum Authority (QCA). Perhaps such research is seen as abstruse and irrelevant to policy and the practice of teaching but its absence throughout the decade of national assessment is surely a national scandal.

Bibliography

- Adams, J. (1912). *The Evolution of Educational Theory* (London, Macmillan).
- AIE (1996). *Assessment in Education*, 3(2).
- Aldrich, R. (1995). *School and Society in Victorian Britain: Joseph Payne and the new world of education* (New York, Garland).
- Aldrich, R. (1996). *Education for the Nation* (London, Cassell).
- Aldrich, R. (1997). *The End of History and the Beginning of Education* (London, Institute of Education).
- Aldrich, V. C. (1963). *Philosophy of Art* (Englewood Cliffs, Prentice-Hall).
- Anderson, R. D. (1995). *Education and the Scottish People* (Oxford, Oxford University Press).
- Arnold, Matthew (1863). *A French Eton*, reprinted in *The Complete Prose Works of Matthew Arnold vol. ii, Democratic Education* ed. R. H. Super (Ann Arbor, University of Michigan Press, 1962), pp 262–325.
- Arnott, M. (1993). Thatcherism in Scotland: an Exploration of Educational Policy in the Secondary Sector (PhD Thesis, Strathclyde University).
- Ayer, A. J. (1946). *Language Truth and Logic* Second edition (London; Penguin).
- Baird, J. (1998). What's in a Name? Experiments with blind marking in A-level Examinations. *Educational Research*, 40(2), 191–202.
- Baird, J. and Jones, B. (1998). Statistical analyses of examination standards: better measures of the unquantifiable? (Associated Examining Board Research Report—RAC/780).
- Bardell, G.; Fearnley, A. and Fowles, D. (1984). *The contribution of graded objectives schemes in Mathematics and French* (Manchester, Joint Matriculation Board).
- Barnes, B. (1974). *Scientific Knowledge and Sociological Theory* (London, Routledge and Kegan Paul).
- Bartholomew, D. J. and Knott, M. (1999). *Latent Variable Models and Factor Analysis* (2nd edition) (London, Arnold).
- Bartholomew, D. J. (1996). *The Statistical Approach to Social Measurement* (San Diego, Academic Press).
- Beardsley, M. C. (1981). *Aesthetics: Problems in the Philosophy of Criticism* (Indianapolis, Hackett).
- Beaton, A. E. and Zwik, R. (1990). *Disentangling the NAEP 1985–86 reading anomaly*. (Princeton, Educational Testing Service).
- Benn, C. and Chitty, C. (1996). *Thirty Years On* (London, David Fulton).
- Berger, P. and Luckmann, T. (1966). *The Social Construction of Reality* (London, Penguin).
- Berry, C. (1997). *Social Theory of the Scottish Enlightenment* (Edinburgh, Edinburgh University Press).
- Best, D. (1985). *Feeling and Reason in the Arts* (London, Allen & Unwin).
- Bierhoff, H. (1996). Laying the foundation of numeracy: a comparison of primary

- school textbooks in Britain, Germany and Switzerland. *Teaching Mathematics and Its Applications*, **15**, 141–60.
- Billington, R. (1988). *Living Philosophy: An Introduction to Moral Thought* (London, Routledge).
- Bourdieu, P. (1989). *La Noblesse d'État: Grandes Écoles et Esprit de Corps* (Paris, Les Editions de Minuit).
- Brock, M.G. and Curthoys, M.C. (1998). (eds.). *The History of the University of Oxford vol. vi, Nineteenth-Century Oxford, Part 1* (Oxford, Clarendon Press).
- Brooks, G. (1997). Trends in standards of literacy in the United Kingdom, 1948–1996 (paper presented at the UK Reading Association conference, University of Manchester, July 1997, and at the British Educational Research Association conference, University of York, September 1997).
- Brown, A., McCrone, D., Paterson, L. and Surridge, P. (1998). *The Scottish Electorate* (London, Macmillan).
- Burnhill, P., Garner, C. and McPherson, A. (1990). Parental education, social class and entry to higher education, 1976–1986. *Journal of the Royal Statistical Society*, series A, **153**, 233–248.
- Burstein, J., Kaplan, R., Wolff, S., and Chi, L. (1997). Using Lexical Semantic Techniques to Classify Free-Responses (Princeton N.J. Educational Testing Service Research Report available on ETSnet at <http://www.ets.org/research/siglex.html>).
- Christie, T. and Forrest, G. M. (1981). *Defining Public Examination Standards* (London, Schools Council/Macmillan).
- Cipolla, C. M. (1969). *Literacy and Development in the West* (London, Penguin).
- Clanchy, M. (1979). *From Memory to Written Record: England 1066–1307* (London, Edward Arnold).
- Collins, R. (1979). *The Credential Society* (New York, Academic Press).
- Committee of Council on Education (1863). *Report of the Committee of Council on Education 1862–63* (London).
- Committee of Council on Education (1872). *Report of the Committee of Council on Education 1871–72* (London).
- Committee of Council on Education (1873). *Report of the Committee of Council on Education 1872–73* (London).
- Committee of Council on Education (1883). *Report of the Committee of Council on Education 1882–83* (London).
- Copeland, J. (1993). *Artificial Intelligence: A Philosophical Introduction* (Oxford, Blackwell).
- Cox, C. B. and Dyson, A. E. (1971). (eds.). *The Black Papers on Education* (London, Davis-Poynter).
- Cresswell, M. J. (1987). Describing Examination Performance: grade criteria in public examinations. *Educational Studies*, **13**(3), 247–65.
- Cresswell, M. J. (1990). Gender Effects in GCSE—Some Initial Analyses (Paper prepared for a Nuffield Seminar at University of London Institute of Education on 29 June 1990) (Unpublished Associated Examining Board Research Report—RAC/517).
- Cresswell, M. J. (1994). Aggregation and Awarding methods for National Curriculum

- Assessments in England and Wales: a comparison of approaches proposed for Key Stages 3 and 4. *Assessment in Education*, 1(1), 45–61.
- Cresswell, M. J. (1995). Technical and Educational Implications of using Public Examinations for Selection to Higher Education. In T. Kellaghan (ed.), *Admission to Higher Education: Issues and Practice* (Dublin, Educational Research Centre and Princeton, International Association for Educational Assessment).
- Cresswell, M. J. (1996). Defining, Setting and Maintaining Standards in Curriculum Embedded Examinations: Judgemental and Statistical Approaches. In H. Goldstein and T. Lewis (eds.), *Assessment: Problems, Developments and Statistical Issues* (London, Wiley).
- Cresswell, M. J. (1997a). *Examining Judgements: Theory and Practice of Awarding Public Examination Grades* (PhD thesis, University of London Institute of Education).
- Cresswell, M. J. (1997b). Can Examination Grade Awarding be Objective and Fair at the Same Time? Another Shot at the Notion of Objective Standards (Unpublished Associated Examining Board Research Report—RAC/733).
- Cresswell, M. J. and Houston, J. G. (1991). Assessment of the National Curriculum—some fundamental considerations. *Educational Review*, 43, 63–78.
- Cressy, D. (1980). *Literacy and the Social Order: reading and writing in Tudor and Stuart England* (Cambridge, Cambridge University Press).
- Damasio, A. R. (1995). *Descartes Error: Emotion, Reason and the Human Brain* (London, Papermac).
- Davis, E. (1993). *Schools and the State* (London, Social Market Foundation).
- Dean, C. (1998). Standards are not parents' top priority. *Times Educational Supplement*, 9 October.
- Dearing, R. (1995). *Review of the 16–19 qualifications* (London, Department of Education).
- Dennett, D. (1993). *Consciousness Explained* (London, Penguin).
- Department for Education and Employment (DfEE). (1997). *Excellence in Schools* (London, Stationery Office).
- Department of Education and Science (1967). *Children and Their Primary Schools. A Report of the Central Advisory Council for Education (England)*. ii (London, DES).
- Devine, M., Hall, J., Mapp, J. and Musselbrook, K. (1996). *Maintaining Standards: Performance at Higher Grade in Biology, English, Geography and Mathematics* (Edinburgh, Scottish Council for Research in Education).
- Devlin, K. (1997). *Goodbye Descartes: The End of Logic and the Search for a New Cosmology of the Mind* (New York, Wiley).
- Dore, R. (1996). *The Diploma Disease*. 2nd edition (London, Institute of Education).
- Dreyfus, H. L. (1992). *What computers still can't do: a critique of artificial reason*. (Cambridge Mass., MIT Press).
- Eagleton, T. (1993). *Literary Theory: An Introduction* (Oxford, Blackwell).
- Eiser, J. R. (1990). *Social Judgement* (Milton Keynes, Open University Press).
- Elwood, J. and Comber, C. (1996). *Gender differences in examinations at 18+* (London, Institute of Education).
- Firestone, W. A. (1998). A Tale of Two Tests: Tensions in Assessment Policy. *Assessment in Education*, 5(2), 175–192.

- Fletcher, S. (1980). *Feminists and Bureaucrats. A study in the development of girls' education in the nineteenth century* (Cambridge, Cambridge University Press).
- Fogelin, R. J. (1967). *Evidence and Meaning: Studies in Analytic Philosophy* (London, Routledge).
- Forrest, G. M. and Orr, L. (1984). *Grade Characteristics in English and Physics* (Manchester, Joint Matriculation Board).
- Foxman, D., Ruddock, G. and McCallum, I. (1990). *APU mathematics monitoring 1984-88 (Phase 2)* (London, Schools Examination and Assessment Council).
- Fremer, J. (1989). Testing Companies, Trends and Policy Issues: A current view from the testing industry. In B. R. Gifford (ed.), *Test Policy and the Politics of Opportunity Allocation: The Workplace and the Law* (Boston, Kluwer).
- French, S., Slater, J. B., Vassiloglou, M. and Willmott, A. S. (1987). *Descriptive and Normative Techniques in Examination Assessment* (Oxford, UODLE).
- Galton, M. (1998). Back to consulting the ORACLE. *Times Educational Supplement*, 3 July.
- Gierl, M. J. and Rogers, W. J. (1996). Factor analysis of the Test Anxiety Inventory using Canadian high school students. *Educational and Psychological Measurement*, **56**, 315-324.
- Goldstein, H. (1983). Measuring Changes in Educational Attainment Over Time: Problems and Possibilities. *Journal of Educational Measurement*, **20**, 369-78.
- Goldstein, H. (1995). *Interpreting International Comparisons of Student Achievement* (Paris, UNESCO).
- Goldstein, H. (1996a) (ed.). *Assessment in Education*, **3**, 2. Special Issue: The IEA Studies.
- Goldstein, H. (1996b). International Comparisons of Student Achievement. In Little and Wolf (1996).
- Goldstein, H. (1999). Performance Indicators in Education. In D. Dorling and S. Simpson (eds.). *Statistics in Society* (London, Arnold).
- Goldstein, H. and Cresswell, M. J. (1996). The comparability of different subjects in public examinations: a theoretical and practical critique. *Oxford Review of Education*, **22**(4), 435-42.
- Goldstein, H. and Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology*, **42**, 139-167.
- Good, F. J. and Cresswell, M. J. (1988a). *Grading the GCSE* (London, Secondary Examinations Council).
- Good, F. J. and Cresswell, M. J. (1988b). *Differentiated Assessment: Grading and Related Issues* (London, Secondary Examinations Council).
- Gould, S.J. (1984). *The Mismeasure of Man* (London, Penguin).
- Gray, J., Hopkins, D., Reynolds, D., Wilcox, B., Farrell, S. and Jesson, D. (1999). *Improving Schools: Performance and Potential* (Milton Keynes, Open University Press).
- Gray, J., McPherson, A. and Raffe, D. (1983). *Reconstructions of Secondary Education* (London, Routledge).
- Green, A., Leney, T. and Wolf, A. (1997). *Convergences and Divergences in European Education and Training Systems* (Brussels, EC Directorate-General XXII (Education, Training and Youth)).

- Green, A., Wolf, A. and Leney, T. (1999). *Convergence and Divergence in European Education and Training Systems* (London, Institute of Education).
- Hacking, I. (1965). *The Logic of Statistical Inference* (Cambridge, Cambridge University Press).
- Hacking, I. (1990). *The Taming of Chance* (Cambridge, Cambridge University Press).
- Hambleton, R. K. and Zaal, J. N. (eds.) (1991). *Advances in Educational and Psychological Testing* (Boston, Kluwer).
- Hargreaves, D. H. (1996). Teaching as a research-based profession: policies and prospects (Teacher Training Agency annual lecture).
- Heath, A. F. and Clifford, P. (1990). Class inequalities in education in the twentieth century. *Journal of the Royal Statistical Society*, series A, **153**, 1–16.
- Holland, P. W. and Rubin, D. B. (1982). *Test Equating* (New York, Academic Press).
- Hollis, M. and Lukes, S. (1982). (eds). *Rationality and Relativism* (Oxford, Blackwell).
- Holmes, E. (1911). *What Is and What Might Be* (London, Constable).
- Jencks, C. (1972). *Inequality: A Reassessment of the Effect of Family and Schooling in America* (New York, Basic Books).
- Johnson, V. E. (1997). An alternative to the traditional GPA for evaluating student performances. *Statistical Science*, **12**, 251–278.
- Kelly, A. (1976). A study of the comparability of external examinations in different subjects. *Research in Education*, **16**, 37–63.
- Kilpatrick, J. and Johansson, B. (1994). Standardised Mathematics Testing in Sweden: The Legacy of Frits Wigforss. *Nordic Studies in Mathematics Education*, **1**, 6–30.
- Koretz, D., Broadfoot, P. and Wolf, A. (1998) (eds.). *Assessment in Education*, **5**(3) (Special Issue on Portfolios and Records of Achievement).
- Kuhn, T. (1970). *The Structure of Scientific Revolutions*, second edition (Chicago, University of Chicago Press).
- Lakatos, I. (1974). *Proofs and Refutations: the Logic of Mathematical Discovery* (Cambridge, Cambridge University Press).
- Little, A. (1996) (ed.). *Assessment in Education*, **4**(1) (Special Issue: The Diploma Disease Twenty Years On).
- Little, A., Wang Gang, and Wolf, A. (1995) (eds.). *Sino-British Perspectives on Educational Assessment* (London, ICRA, Institute of Education).
- Little, A. and Wolf, A. (1996) (eds.). *Assessment in Transition: Learning, monitoring and selection in international perspective* (Oxford, Pergamon).
- Long, H. A. (1985). Experience of the Scottish Examinations Board in developing a grade-related criteria system of awards (Paper presented at the 11th annual conference of the International Association for Educational Assessment held in Oxford, England).
- Macaulay, Lord (1898). *Collected Works*, 12 vols. (London, Longmans Green).
- Mackenzie, D. A. (1981). *Statistics in Britain 1865–1930. The Social Construction of Scientific Knowledge* (Edinburgh, Edinburgh University Press).
- McKenzie, D. (1994). The irony of educational review. *New Zealand Annual Review of Education*, **4**, 247–59.
- McLean, L. D. (1996). Large-Scale Assessment Programmes in Different Countries

- and International Comparisons. In H. Goldstein and T. Lewis (eds.), *Assessment: Problems, Developments and Statistical Issues* (Chichester, Wiley).
- McPherson, A. and Willms, J. D. (1987). Equalisation and improvement: some effects of comprehensive reorganisation in Scotland. *Sociology*, **21**, 509–39.
- Madaus, G. and Raczek, A. (1996). Turning Point for Assessment: Reform Movements in the United States. In Little and Wolf (1996).
- Menet, J. (1874). *A Letter to a Friend on the Standards of the New Code of the Education Department* (London, Rivingtons).
- Morrison, H. G., Busch, J. C. and D'arcy, J. (1994). Setting reliable national curriculum standards: a guide to the Angoff procedure. *Assessment in Education*, **1**, 181–199.
- Murphy, R. J. L. (1982). Sex differences in Objective Test performance. *British Journal of Educational Psychology*, **52**, 213–19.
- Murphy, R. J. L., Burke, P., Cotton, T., Hancock, J., Partington, J., Robinson, C., Tolley, H., Wilmot, J. and Gower, R. (1996). *The Dynamics of GCSE Awarding: Report of a project conducted for the School Curriculum and Assessment Authority* (London, SCAA).
- Newcastle Report (1861). *Report of the Commissioners appointed to inquire into the State of Popular Education in England*, PP 1861 XXI (ii) (London).
- Newton, P. (1996). The reliability of marking of GCSE scripts: Mathematics and English. *British Educational Research Journal*, **22**, 405–20.
- Newton, P. (1997a). 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. (1997b). Examining Standards Over Time. *Research Papers in Education*, **12**(3), 227–48.
- Orr, L. and Forrest, G. M. (1984). *Investigation into the relationship between grades and assessment objectives in History and English examinations* (Manchester, Joint Matriculation Board).
- Orr, L. and Nuttall, D. L. (1983). *Determining Standards in the Proposed Single System of Examinations at 16+* (London, Schools Council).
- Paterson, L. (1992). The influence of opportunity on aspirations among prospective university entrants from Scottish schools, 1970–1988. *Statistics in Society, Journal of the Royal Statistical Society, series A*, **155**, 37–60.
- Paterson, L. (1995). Social origins of under-achievement among school-leavers. In L. Dawtrey, J. Holland, M. Hammer and S. Sheldon (eds.), *Equality and Inequality in Education Policy* (Milton Keynes, Open University Press).
- Paterson, L. (1997). Student achievement and educational change in Scotland, 1980–1995. *Scottish Educational Review*, **29**, 10–19.
- Paterson, L. (1998). The Scottish parliament and Scottish civil society: which side will education be on? *Political Quarterly*, **69**, 224–33.
- Paterson, L. (forthcoming). Scottish traditions in education. In H. Holmes (ed.), *Compendium of Scottish Ethnology, vol. 11* (Edinburgh, Scottish Ethnological Research Centre).
- Paterson, L. and Raffe, D. (1995). Staying on in full-time education in Scotland. *Oxford Review of Education*, **21**, 3–23.
- Payne, J. (1872). 'Why are the Results of our Primary Instruction so Unsatisfac-

- tory?', *Transactions of the National Association for the Promotion of Social Science*.
- Phillips, M. (1996). *All Must Have Prizes* (London, Little, Brown and Company).
- Pirsig, R. M. (1974). *Zen and the Art of Motorcycle Maintenance: An Inquiry into Values* (London, Bodley Head).
- Plewis, I. (1998). Inequalities, Targets and Zones. *New Economy*, 5, 104–8.
- Plewis, I. (1999). What's Worth Comparing in Education? In D. Dorling and S. Simpson (eds.). *Statistics in Society* (London, Arnold), 273–80.
- Pole, D. (1961). *Conditions of Rational Inquiry: A Study in the Philosophy of Value* (London, Athlone).
- Power, M. (1997). *The Audit Society: Rituals of Verification* (Oxford, Oxford University Press).
- QCA (1998). *GCSE and GCE A/AS code of practice* (London, Qualifications and Curriculum Authority).
- Reynolds, D., Creemers, B. P. M., Stringfield, S. and Teddlie, C. (1998). Climbing an educational mountain: conducting the International School Effectiveness Research Project. In G. Walford, *Doing research about education* (Lewes, Falmer Press).
- Roach, J. P. C. (1971). *Public Examinations in England 1850–1900* (Cambridge, Cambridge University Press).
- Robertson, C. (1992). Routes to higher education in Scotland. *Scottish Educational Review*, 24: 3–16.
- Rose, S. (1997). *Lifelines, Biology, Freedom, Determinism* (London, Penguin).
- Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education*, 13, 191–209.
- Schools Council (1979). *Standards in Public Examinations: Problems and Possibilities*, Report from the Schools Council Forum on Comparability (London, Schools Council).
- SEC (1984). *The development of Grade-related Criteria for the General Certificate of Secondary Education—a briefing paper for working parties* (London, Secondary Examinations Council).
- SEC (1985). *Reports of the Grade-related Criteria Working Parties* (London, Secondary Examinations Council).
- SEC (1986). Draft Grade Criteria. *SEC News Number 2* (London, Secondary Examinations Council).
- SEC (1987). Grade Criteria—Progress Report. *SEC News Number 6* (London, Secondary Examinations Council).
- Shavit, Y. and Blossfeld, H. P. (1993). *Persistent Inequality: Changing Educational Attainment in Thirteen Countries* (Boulder, Col., Westview Press).
- Skolöverstyrelsen (1980). Quoted in J. Kilpatrick and B. Johansson (1994). Standardised Mathematics Testing in Sweden: The legacy of Frits Wigforss. *Nordic Studies in Mathematics Education*, 1, 6–30.
- Smith, J. V. and Hamilton, D. (1980) (eds). *The Meritocratic Intellect* (Aberdeen, Aberdeen University Press).
- Start, B. and Wells, K. (1972). *The trend of reading standards* (Slough, National Foundation for Educational Research).
- Stedman, L. C. (1998). An Assessment of the Contemporary Debate over US

- Achievement. In D. Ravitch (ed.), *Brookings Papers on Education Policy* (Washington DC, Brookings Institution Press), 53–119.
- Stephens, W. B. (1987). *Education, Literacy and Society, 1830–70: the geography of diversity in provincial England* (Manchester, Manchester University Press).
- Sutherland, G. (1973a). *Policy-Making in Elementary Education 1870–1895* (Oxford, Clarendon Press).
- Sutherland, G. (1973b) (ed.). *Matthew Arnold on Education* (London, Penguin).
- Sutherland, G. (1984). *Ability, Merit and Measurement. Mental testing and English education 1880–1940* (Oxford, Clarendon Press).
- The Scotsman Education* (1998). 30 September: 4–5.
- Thom, D. (1986). The 1944 Education Act: the ‘art of the possible. In Harold L. Smith (ed.), *War and Social Change: British Society in the Second World War* (Manchester, Manchester University Press), 101–28.
- Vincent, D. (1989). *Literacy and Popular Culture: England 1750–1914* (Cambridge, Cambridge University Press).
- Walden, G. (1996). *We Should Know Better: solving the educational crisis* (London, Fourth Estate).
- Wang Binhua (1995). Comparing HSCE in the People’s Republic of China and GCSE in England. In Little, Wolf and Wang Gang (1995).
- Wang Gang (1995). The Development of Public Educational Examinations in China from 1980. in Little, Wolf and Wang Gang (1995).
- William, D. (1996a). Meanings and Consequences in Standard Setting. *Assessment in Education*, 3(3), 287–307.
- William, D. (1996b). Standards in examinations: a matter of trust? *The Curriculum Journal*, 7(3), 293–306.
- Wilmot, J. and Rose, J. (1989). *The Modular TVEI Scheme in Somerset: its concept, delivery and administration* (Report to the Training Agency of the Department of Employment, London).
- Wolf, A. (1995). *Competence Based Assessment* (Buckingham, Open University Press).
- Wolf, A. and Steedman, H. (1998). Basic Competence in Mathematics: Swedish and English 16 year olds. *Comparative Education*, 34, 3.
- Wood, R. (1991). *Assessment and Testing: A survey of research* (Cambridge, Cambridge University Press).
- Young, M. (1958). *The Rise of the Meritocracy 1870–2033* (London, Penguin).