120

EQUITY IN TESTING AFTER GOLDEN RULE

by

Harvey Goldstein
Institute of Education
University of London
London
England

Paper read to American Educational Research Association meeting,
San Francisco, March 27-31, 1989

1 INTRODUCTION

Recent articles in the journal 'Educational Measurement: Issues and Practice' (in Summer 1987, and Spring 1988) have explored issues arising from the debate between the Golden Rule Insurance Company (GRIC) and Educational Testing Service (ETS). The debate has centered around the procedures originally agreed between GRIC and ETS for minimising Black/White test score differences. These procedures were based upon choosing those test items which showed the smallest group differences after various standard item screening techniques had been employed to yield candidate items for inclusion in the test.

Among the important political and social issues which this debate has highlighted, is that of the relationship between the technical characteristics of a test and its social impact. This relationship, however, is only explored partly in the above articles and the purpose of the present paper is to extend the debate by questioning whether the 'bias elimination' procedures discussed by Linn and Drasgow(1987) and Anrig(1988) really address the point at issue. First, however, some historical perspective may be useful.

2 HISTORICAL PRECEDENCE

It seems clear that one of the major motivations for introducing large scale testing, both for example in the U.K. During the 19 Th century and the U.S.A. In the early 20th century, was a concern with equity and an attempt to select on merit. At the same time, as Gould (1981) demonstrates, cultural expectations played an important role in the way in which the tests functioned, and in particular the patterns of differences between groups of the population. As Weiss(1987) points out, it is possible to construct tests, by selecting appropriate item contexts, to reduce or reverse Black/White differences. Goldstein (1986) makes a similar point in relation to gender differences. Thus, in part at least, observed group differences may be have as much to do with the cultural expectations of test designers as with any 'real' differences in knowledge or ability.

One of the ways in which new tests are validated is to compare the performance of their new items against items from an existing, often well established, test. It is not too difficult to see how such a procedure can become an effective vehicle for the perpetuation of old cultural expectations about how items should discriminate between groups. In recent times, such an effect has probably been mitigated by the attention to item content and a sensitivity to ethnic and sexual stereotyping. Nevertheless, ethnic and gender response differences remain (this is what the Golden Rule dispute is all about) and a key question is whether such differences are 'artifacts' of the items or in some sense 'real'. The next section looks at this issue in detail.

3 THE REALITY OF GROUP DIFFERENCES

In the article by Linn and Drasgow, and that by Anrig, there is an implicit assumption that a measure is available for the true 'skill' or 'ability' a test item is supposed to measure. Anrig, for example, says that subjects who *'know the same amount about a test item'* should have a similar chance of answering it correctly *'regardless of their race, sex, or ethnic background'*. Linn and Drasgow state that an adequate approach to detecting item bias *'requires a means of distinguishing between differences that are due to group differences in the developed skills of the test takers and those that are due to extraneous factors'*. The problem, of course, is to measure in some suitable way these 'skill' or 'knowledge' factors. In practice, test constructors use the total set of items available (or some equivalent test) to provide such measures, and then to identify unusual or 'outlying' items as candidates for possible bias.

Linn and Drasgow propose a 1-dimensional item response model criterion whereby an item is judged to be unbiassed if its characteristic or response curve is the same in each group except for a shift in location along the 'ability' scale. Unfortunately, any difference along the ability scale can be interpreted either as a 'real' difference between groups, or as a between group 'bias'. Since the ability scale is estimated effectively as a weighted average function of the item responses in each population, such a procedure simply detects 'outlying' items. Thus, for example, if all the test items show a similar group difference, it is a matter of judgement whether we wish to interpret this as a biassed test or a real group difference or some mixture of these two. The point is that there is no agreed *external* criterion for making a judgement, and no amount of statistical modelling can avoid that fact. One of the problems with Item Response Theory (for 'theory' read 'models') is that its mathematical complexity masks its logical inadequacy.

4 THE LEGITIMACY OF GOLDEN RULE PROCEDURES

If it is accepted that 'technical' approaches to detection of item bias are inadequate because they are essentially tautological, we are left with the best efforts of test constructors and the elimination of content bias while attempting to maintain educational or psychological relevance. Some writers on this topic (see for example, Humphreys, 1986), base their recommendations on an assumption about the existence of a valid outcome criterion against which group test differences can be judged. Needless to say, this begs the question of how such a criterion is to be found. If bias exists then it will have affected all the criteria which could be judged relevant. Such an approach hardly seems helpful.

At the end of the day, as the Golden Rule controversy highlights, the test constructor is left with candidate test items which are technically similar and also adequate, for example by having reasonable within-group discriminations. Nevertheless, some will show smaller and perhaps reversed group differences than others. Naturally, as several writers have pointed out, a mechanical application of a formula, such as that agreed between ETS and GRIC, can lead to bad judgements. Nevertheless, a requirement to select those items which minimize group differences on the final test, does appear to meet a general requirement for equity. Bond (1987) takes a similar view. Indeed, given the historical circumstances of test construction, we might also require test constructors to search and develop systematically items which showed opposite effects to those expected.

Interestingly, Anrig's (1988) reply to Rooney (1987) does not question such a procedure in general terms. Indeed, it is difficult to see that ETS or any other test agency could so do since, by definition, the items which are available for consideration, after standard vetting procedures have been carried out, are to all intents and purposes technically equivalent; except that they exhibit group differences.

5 CONCLUSIONS

Having argued in favour of the general principle of a procedure like that of the Golden Rule, I am left with two outstanding questions.

The first concerns how we might operate such a procedure, both politically and technically. I do not propose to discuss this in detail here, save to point out that if it is to be taken seriously, it is not a matter which should be left to the testing agencies. It is a matter of social concern which goes beyond those agencies and the testing profession too: at the least it should involve educational professionals drawn from diverse backgrounds, including those whose major concern is other than testing.

The second question concerns the attitude of much of the testing profession itself. The reaction to Golden Rule, at least as expressed in the articles referred to, largely has been to retire into technicalities. Thus Linn and Drasgow claim that '*the most widely accepted psychometric approach to this problem* (of item bias) *is based on Item Response Theory*'. Widely accepted by whom one may ask? Likewise, Anrig refers to '*the theoretical and analytical sophistication of the ETS methodology*'. It might not be innapropriate to recall an older use of the term 'sophistication', namely '*the process of investing with specious fallacies or of misleading by means of these*' (Oxford English Dictionary).

It seems that the testing profession needs to develop a greater awareness of the social and political implications of its techniques. It also needs to display a greater willingness to expose the logical structures of its techniques, rather than reverting to mathematicisation, when challenged. If it cannot undertake such tasks itself, then it should not be too surprised if others seek to do this for it.

6 SUMMARY

The dispute between ETS and the Golden Rule Insurance Company raises important social issues. It is argued that the testing profession needs to recognise that there may be no purely technical solutions to problems of test bias and that the process of test construction and analysis should recognise the need to make ideological and social choices. In particular it is argued that the use of item response models to attempt to resolve problems of test bias, is both innapropriate and misleading.

7 POSTSCRIPT

This paper was submitted for publication in the journal 'Educational Measurement: Issues and Practice'. It was sent to three referees, all of whom recommended outright rejection! Their detailed responses neatly illustrate many of the arguments of the paper itself, and are worth a brief summary.

All three referees claim that there is nothing original in the paper; a somewhat curious reason for rejection. If a critique is appropriate, then the fact that it draws upon earlier arguments hardly seems germane. More importantly, the general attitude of the referees is to deny that any fundamental problem exists.

Thus, one referee claims that '*good*' tests show no prediction bias against minorities and furthermore that tests which reduce Black-White differences do not predict well. She or he in effect is unwilling to recognise that there is a problem and concludes with the somewhat condescending tautology that '*disadvantaged people really are disadvantaged*'.

Another referee claims never to have been aware that the cultural expectations of test developers have resulted in detriment to any group. This of course precisely illustrates my point. The cultural conditioning of a person does not allow easily that person to observe the effect that the culture itself is exerting. Somewhat revealingly, this referee takes me to task for stating that item response models are mathematically complicated. '*Actually*', she or he says, '*its very simplicity....is one of its claims to elegance*', and goes on to argue that such models '*may best serve where disagreement exists with respect to qualitative criteria for making decisions regarding bias*'. I think it would not be too much of a distortion to summarise this stance as 'the model is too elegant not to be true'. Needless to say, one person's elegance may be another person's oversimplification.

Finally, this same referee echoes what I believe is a typical response of the testing profession. He or she refers to good test development practice lying in the adherence to '*test specifications, a blueprint prepared by subject matter experts who are knowledgeable about the purpose of the test and the characteristics of the intended test-taking population.*' This effectively closes the circle around the profession. The real issue, however, remains. Namely, how long the profession can continue to ignore the claims of those who do not wish to share all, or even some, of its basic assumptions and beliefs.

8 REFERENCES

Anrig, G.R. (1988) ETS replies to Golden Rule on "Golden Rule". Educ. Meas.: Issues and Practice 7:20-21.

Bond, L.(1987). The Golden Rule settlement: A minority perspective. Educ. Meas.: Issues and Practice 6:18-20.

Gould, S.J. (1981) The Mismeasure of Man, New York:W. W. Norton.

Goldstein, H. (1986) Gender bias and test norms in educational selection. Research Intelligence (British Educational Research Association Newsletter): May, 2-4

Humphreys, L. (1986) An analysis and evaluation of test and item bias in the prediction context. J. Applied Psychology, 71, 327-33.

Linn, R. L., And Drasgow, F. (1987) Implications of the Golden Rule settlement for test construction. Educ. Meas. : Issues and Practice, 6: 13-17

Weiss, J. (1987) The Golden Rule bias reduction principle: a practical reform. Educ Meas. : Issues and Practice, 6: 23-25.