3c

THE DEMOCRATISATION OF TEST CONSTRUCTION: A RESPONSE TO THE PROBLEMS OF EDUCATIONAL MEASUREMENT IN A MULTI-ETHNIC SOCIETY

G. K. VERMA
UNIVERSITY OF BRADFORD

S. KEMMIS AND B. MACDONALD
UNIVERSITY OF EAST ANGLIA

PAPER PRESENTED AT THE
INTERNATIONAL SYMPOSIUM
ON EDUCATIONAL TESTING,
UNIVERSITY OF LEYDEN
THE NETHERLANDS

JUNE 27 - 30, 1977

The Democratisation of Test Construction: A Response to the Problems of Educational Measurement in a Multi-ethnic Society.

Introduction:

Friedenberg (1969) has stated succinctly the general problem that we seek to address in this paper:

> "Educational measurement is an inherently conservative function, since it depends on the application of established norms to the selection of candidates for positions within the existing social structure on terms and for purposes set by that structure. It cannot usually muster either the imagination or the sponsorship needed to search out and legitimate new conceptions of excellence which might threaten the hegemony of existing elites.

The particular problem that concerns us, ethnocentric bias in psychometric tests, has received little attention so far in Britain, but this is a situation we expect to change rapidly in the light of contemporary political and educational trends. Two of these trends deserve a brief mention before we proceed to a scrutiny of the relevant testing issues.

Although there are countervailing indices, it would be hard to resist the proposition that British social policy has become markedly racist over the past two decades, a trend reflected not only in legislation governing immigration, but in the changing rhetoric of political party spokesmen. Labour Party views in particular have moved steadily to the right in pursuit of electoral safety, a process convincingly documented by Moore (1975). Party differences have narrowed as they compete for a kind of middle ground stance which Downing and Schlesinger(1976) scathingly term "reasonable racism".

The extent to which this political manoeuvring represents an accurate reading of popular sentiment, or the extent to which it actually creates the sentiment it seeks to reflect, is a matter for conjecture. Britain is only beginning, it seems, to come to terms with the fact of multi-ethnicity and cultural diversity; perceptual and attitudinal ambiguities are part of the legacy of a recent history of population change which has yet to be fully assimilated.

The second trend, unrelated to racism except in so far as both can be seen in part as responses to economic failure, is the more recent renaissance in large scale educational messurement under government sponsorship. The creation in 1974 of an Assessment of Performance Unit within the Department of Education and Science, to mount a national testing operation on the school population in every area of the curriculum, has provided a fresh impetus to a testing industry that was in a state of some decline. Although at this point in time no one can be sure to what uses the many tests now under development will be put, it does seem certain that the role of testing in the education service will be significantly enhanced in the near future.

If Friedenberg's analysis, with which we concur, is correct, there is a very serious question to be raised about the extent to which the planned expansion of centrally developed; nationally applied tests is likely to discriminate against ethnic minority groups in Britain, or at the least to produce misleading information for policy determination.

The Defects of Multi-ethnic Testing

Measurement, assessment or testing - whichever words we decide to use - has been the target of much criticism in the last ten years. There are many - both within and outside the social sciences - who frequently express their scepticism about both the process of quantification and its interpretation. This applies with particular force to the area of race research or assessment in multi-ethnic societies where testing is viewed as a dubious activity. Even the testing community concede that concern, is not unjustified or trivial. The early testing movement certainly reflected the strong elitist and racist values of the testers, and the common elements of their culture. For example, Thorndike's (1920) work, 'The Psychology of the Half-Educated Man' is a clear indication of how testers measured people against their own model of a successful individual.

In the field of Western education the testing movement, despite its commitment to meritocratic ideals, produced massive discrimination between various racial groups (Karier, 1973). Tests were utilized not only to discriminate against children in their education but to restrict employment opportunities for minority groups. We know from our experience that in multi-racial work situations the use of objective tests as screening devices for applicants is advanced by some employers

as arguments for their well-intentioned lack of bias in appointment and promotion procedures. However, the fact is that there is no guarantee of such a lack of bias in these selections. (Sidney Irvine, 1973). Furthermore, this movement helped to develop and perpetuate the myth of 'scientific objectivity'. Even today, most intelligence tests utilized in schools and other walks of life in America and Britain clearly reflect the common elements of a particular culture. Some psychologists have attempted to introduce broad culturally based tests while others claim to have sought the impossible: culture free or culture-fair intelligence tests.

In recent years the most striking aspect of the controversy in this field was the value attached to the 1969 Jensen report on differences in intelligence between blacks and whites. There has been a great deal of discussion on the reliability of the methods used in collecting the evidence. As soon as the Jensen report appeared it became the object of vigorous criticism, both for its methodological shortcomings and for technical inadequacies in sampling and in procedures. Most current criticisms of testing amount to statements that tests are invalid or biased, that the use of tests is a cold, machine-like process, and that the results of the tests are often misused.

Some may be tempted to ask why on earth should we bother to measure people's characteristics, but this is to suggest that since there are some reckless drivers on the road, driving should be banned. A major attraction of measurement is that it holds out the possibility of a more precise appraisal of human characteristics which can be used in a variety of decision-making activities. Churchman, (1971) is perhaps right in suggesting that measurement should be 'a decision-making activity designed to accomplish an objective'. From our perspective, measuring tools can make useful contributions in the analysis of certain types of social phenomenon if they ure used cautiously, creatively and in an egalitarian way.

Although far more sophistication has been introduced in the process of measurement in the last decade, some of the issues have remained unresolved.

There may well be a case for saying that we tend to rush into testing populations with more enthusiasm than care. For example, tests of intelligence, aptitude, attitude and personality are constructed and standardized for one particular ethnic group but are used for other

ethnic and racial groups. Generalisations from the results of these tests have often given rise to a wide variety of misleading interpretations, especially from those who have little understanding of the populations. Thus, in the absence of adequate validation, tests tend to develop into self-fulfilling prophecies.

As Karier's work in particular indicates, many widely used tests are thoroughly permeated by the cultural and ideological perspectives of their developers. It should not be thought that this cultural and ideological influence is "contamination" in the sense that the tests were generally acceptable but blemished by the cultural perspectives of their developers - the tests are the products of their culture and ideology. The tests themselves are cultural artefacts, not merely bent at the edges by cultural biases.

This kind of critique has recently been levelled at policy-oriented international research also. The I.E.A. study of comparative cognitive achievement in twenty countries drew the following comments from William Platt (1975) of UNESCO:

> "There is good reason for suspecting that the tests inadvertently were not culturally fair, that they were overdependent upon reading ability, upon Western concepts and values, and upon experience with the multiple choice format."

The issues are no longer matters of mainly academic debate. The controversy about test bias, particularly in intelligence and attainment testing, has led to the suspension of testing in many parts of America. In 1969 the American Association of Black Psychologists expressed its concern about test bias by calling for a moratorium on all testing of black people "until more equitable tests are available". Although this call for a maratorium has been contested by some on the grounds that the absence of normative checks would result in increased discrimination, it has had considerable political success, and has influenced test developers to shift their attention to measures of inter-ethnic attitudes and perceptions.

But measuring racial attitudes has proved to be as problematic technically as measuring race ability and achievement had become politically. There are only a few attitude measures and those are technically defective and are difficult to adapt forparticular populations. Technical deficiencies are numerous; they include lack of standardization, poor statistical precision and thin empirical grounding, crudity, lack of credibility and obsolescence.

Above all, the ethnocentrism of attitude instruments is a well-known and persistent problem that seems incapable of resolution. Problems of ethnocentric bias have been recognized by many researchers in the field of race research (Biesheuval, 1949; Anastasi, 1959; Schwarz, 1962), yet ethnically biased tests are often used in multi-ethnic situations, simply because test use is running ahead of test development. In the British context, the only satisfactory tests of attitude in this area (Husband, 1972; Warr, 1967) have been constructed on an ethnocentric basis, yielding scores of attitude to other races which are valid for one race only i.e. in general, Western whites, the race to which the test developers belong. Such tests contain culturally embedded assumptions which unreliably estimate or discriminate against cultural minorities. This sort of test seems to have credibility and political acceptability as a mono-ethnic measure, but its validity for measuring the attitudes of ethnically mixed populations, and particularly of migrant populations such as obtain in the U.K., is highly questionable.

In view of the inherent defects in most direct methods of attitude measurement, some test users have turned to indirect measures of inter-group relations, they have more satisfactory psychometric characteristics, but their indirectness has proved to be of a low general acceptability in multi-ethnic situations.

Given the various defects associated with existing tests, there is a strong case for saying that the instruments in use in the multi-ethnic context are not finely calibrated tools but, at best, rough-and-ready devices in a primitive stage of development. Testing in the area of race relations presents a problem unlikely to be solved immediately and conclusively.

The Need for Testing

If our analysis so far is correct, then we must draw the conclusion that the current controversy about tests in the area of race, whether intellectual or attitudinal tests are being considered, is a political and ideological one. It is political in the sense that the use of tests can be perceived by those tested, and by some measurement specialists, as discriminatory insofar as it can differentially affect the life-chances of members of different ethnic groups. It is ideological in the sense that it embodies values which can turn out to have potentially discrimanatory consequences. The defence of the status quo in educational testing, it might be argued, is thus

the defence of a psychometric-scientific ideology which is politically conservative - the views of Karier and Friedenberg quoted above would certainly support such an argument. This scientific ideology depends on the premise of the uniformity of nature (Hamilton, in press) - in psychometrics rendered as the uniformity of the nature of intelligence or the structure of attitudes - and its political counterpart is meritocratic universalism.

To counter the legitimate attacks now being made on testing, it is necessary to step outside the uniformist, universalistic framework and reconsider the nature of tests themselves. Shortly, we would like to propose an alternative approach to testing - one of many, perhaps, and at best only a conjecture about possible future developments - but one which we believe merits attention in the light of the current controversy.

We should not be too complacent about the need for such alternatives, however: without new approaches, the controversy will continue and ultimately will lead to such a decline in confidence in testing as to place its future in serious doubt. At the outset, a defence must be given of the need for testing in an area of such intense social sensitivity.

We do not believe that testing in the area of race should be abandoned. But, in this area in particular, we are in agreement with the majority of psychometricians who argue that tests must be used with far more care than they have been. It is up to test developers to ensure that tests can be used properly, that their framing assumptions are made explicit to users, and that sufficient information is provided on the purposes for which the test is appropriate and on the settings for which it was designed. Test developers have an increasing responsibility to see that tests are used within their design limits and that lay use of test results is informed by specialist advice.

The central problem is that a society determined to improve race relations needs to be aware of the policies which best promote changes and those which are counter-productive. This implies repeated and systematic evaluation of these policies.

In Britian, for example, despite the political shift to the "right" referred to earlier, there is considerable investment to improve relations between various ethnic groups. Current policy is controversial i.e. people of different ethnic/cultural groups disagree about its likely effects and about the 'best intentions' of those who help to shape it.

However, these efforts need to be evaluated in order to guide future policy. It is argued by some that psychometric measures of the effectiveness of policy initiatives may have a greater degree of political and social acceptablity than alternative forms of assessment. Given the degree of suspicion which exists between ethnic/cultural groups which fear deprivation of social and educational opportunity there is a need for rethinking he strategies of test construction in a more democratic way. Testing may be les vulnerable to criticism on the grounds of bias than other methods of assessment, but only if an approach to testing can be developed that reflects rather than denies the legitimate diversity of social democratic pluralism.

An Alternative Approach

The sort of confusions rife in the field of race research suggest that there is still a pressing need for a valid, simple and flexible instrument for measuring inter-ethnic attitudes. Such an instrument must be acceptable to minority groups in a multi-cultural, multi-ethnic society. In this section we would like to propose an alternative approach to the construction of tests in the area of race, arguing from the two aspects of the current controversy over testing identified earlier. First we will argue from a scientific-ideological perspective (against uniformism) and then from a political perspective (against meritocratic universalism). The arguments apply more widely, but it will be useful to take a specific case to demonstrate the feasibility of the approach.

In the past, the construction of a test like one of inter-ethnic attitudes has been predicated on the questionable assumption that the instrument itself would be capable of rendering inter-group differences in a non-controversial, "neutral" technical language (e.g. reports of group means and standard deviations, regression functions, discriminant functions). Differences between groups appear as differences in patterns of response to the standard instrument.

In general, this view follows from the psychometric ideology previously referred to , that of the uniformity of the nature of the phenomena across contexts, with differences in the manifestation of the phenomena being attributable to contextual differences. This view assumes that the phenomenon itself (e.g. intelligent performance) is uniform. Under this view the assumption is that intelligence is a kind of "humour of the mind", existing as a substance or like height which some people have more of than others. Such a view entitles us to use a standard stimulus

(the instrument) which, through the mediation notion of test validity, bears a specified relation to the phenomenon. Leaving aside the extreme operationalist view that "intelligence is what the test measures" in which the theory of validation is taken as entirely unproblematic, we can see that for any more complex view of test validity the validation operation ends up defining the phenonemon through the relation of the test to the criterion. But validity is conferred on the instrument, it does not inhere in it; so it becomes reasonable to ask for whom the test is a valid meauure.

Classicall y, test developers have stressed the need to develop local norms for an instrument. In this spirit, Messick and Anderson (1974) have recently reminded us that is is possible to use within-group test validity to secure fair test use for multi-racial populations. Similarly, Cronbach (1971) argues for local validation of tests based on local needs and local prediction criteria: The test-criterion relationsihp will be affected by local circumstnnces of testing, interpretation and use, so local validation is necessary to maximise test usefulness.

But it is possible to take the argument a stage further. In the context of aptitude-treatment interactions, Cronbach (1975) has stressed the need for researchers to attend to unique local circumstances and contextual effects which do not merely affect the manifestation of aptitude-treatment interactions, but which fundamentally affect the nature of the variables themselves. Aptitude and treatment are no longer formally "variables"; they are in turn a complex epression of context. Analogously, the alternative approach we want to propose begins by questioning the assumption that the instrument should be a standard stimulus. Instead, we want to argue for local development of tests.

Given local criteria for test use (in selection, attitude description, or policy assessment), it should be possible to develop tests which best serve their purposes because entirely adpated to local criteria and circumstances. Furthermore, we believe that the notion of "local" that we are employing here does not only refer to geographical location, but also to cultural or ethnic location. The problem is one of defining the boundaries of commonality. Take a Glasgow Pakistani: is he more like other Glaswegians? Other Pakistanis? Others for whom English is a second language? Deciding questions of cultural and ethinic location is a matter for further research.

When we challenge the uniformist view that the phenomenon is the same across groups and contexts of application, we find ourselves led to a view of testing which does not depend on the notion that the test must be a standard stimulus. Local test development will lead to non-standard tests with common purpose: in this case different tests, constructed by and validated for different ethnic groups. Though it is not the purpose of this paper to argue beyond this case, it would seem applicable in principle to the development of different tests of intelligence also; the central premise being that intelligence is a culturally-located phenomenon, and is thus structurally and qualitatively different in different ethnic and cultural circumstances.

Now the foregoing argument, overturning as it does a long- standing assumption of educational measurement (one whose roots can be traced back to Galton and his notion of "natural ability"), will seem to some sufficiently controversial as to constitute meagre grounds for changing our approach to the development of tests of inter-ethnic attitudes. A second and independent line of arguemt, based on the acceptability of such tests in assessing social policy, may be constructed.

If social policy is to be informed by tests then in a democratic society it must be demonstrated that the tests are fair to the groups being measured. So long as such tests are defective in the variety of ways considered earlier then they will lack credibiiity to the groups whose attributes they purport to describe. Lacking credibility, the tests will lack acceptability; lacking acceptability, they cannot hope to inform social policy. Consent to be governed in a democratic society depends upon consensual agreement about the justice of the procedures of government (hence political prisoners reject the authority of the courts); only if agreement can be secured as to the validity of tests of inter-ethnic attitude for the groups tested can such tests be acceptable as informing social policy.

In a multi-cultural, multi-ethnic society, a reasonable way forward would thus be to adopt an ethnocentric basis for the developemnt of a set of loosely parallel forms of instrument to serve as a measure of inter-ethnic relations. Each scale should be developed within the boundaries and under the control of the ethnic community concerned. In the development of such tests it would be essential to gain an understanding of the expectations, habits, norms and values of the groups for whom the test is being designed.

The characteristics of racial, ethnic and migrant groups diff r widely and qualitatively within each society, between societies, and at different historical periods. In studying race relations it is therefore necessary to come to terms with the qualitative differences between sets of inter-ethnic relations. From the ethnic and cultural differences between groups it follows that inter-ethnic perceptions will vary depending upon the group perceiving and the group being perceived. To measure inter-ethnic perception, therefore, a set of measures rather than a single measure must be used.

Initially, a test development programme in the area of inter-ethnic attitudes might begin with a scale which was not ethnically-based, but by involving members of different ethnic groups in the development of tests appropriate and acceptable to themselves, it would be possible to move rapidly towards the development of a set of ethnically-based instruments. Once the ethnocentric base of each of the tests is established (both conceptually and psychometrically), it may be possible to move from monitoring inter-ethnic perception/reaction to a situation where such instruments could help in defining a language for negotiation between groups. That is, by using each group's own test of its perceptions of other groups as a basis, it may be possible to characterise the differences in criteria, priorities and attitudinal structures between groups in terms of the instruments themselves: the instruments will have become expressions of the ethnic and cultural perspectives of their "owners".

It would be a mistake to think that a set of ethnically-based instruments like this will define differences in perspectives, however, Each test, as a cultural artefact, will reflect its embeddedness in the cultural and ethnic perspectives of its "owners" but this does not imply that a description of differences between the tests will be an accurate description of differences between the cultures. Such a description would run into the familiar problem of language that, being a medium of communication it appears to cross the cultural divides between groups without change of meaning. As every translator knows, however, different language groups have different perspectives created and maintained in language; the imposition of a common language does not eliminate those differences. Thus, a description of differences between ethnically-based tests does not provide a new "neutral" metalanguage: each group may want to describe the differences between its own and other groups' tests in its own way. Nevertheless, by discussing (not defining such differences, it may be possible to increase inter-ethnic inderstanding through a programme of ethnically-based test development, and to secure agreement about

the acceptability of such tests because each test is acceptable to its own group.

We believe that a test development programme such as this, by exposing the cultural bases of inter-ethnic perception and by helping to identify the shared world-views of members of particular ethnic groups, offers new possibilities in the formulation and evaluation of social policy, particularly policy affecting minority groups. It may provide a mechanism for gathering perspectives on policies and their consequences from within identifiable groups which may be differentially advantaged by them. For example, evaluators of curriculum development, in the area of race relations may want to describe changes in inter-ethnic attitudes as a consequence of teaching and learning. Given ethnically-based tests of inter-ethnic perception, they will be able to describe more precisely how students have responded, and in particular, how the attitudes of different ethnic groups have changed in their own terms, rather than in terms of some general instrument which may be culturally-insensitive in terms of the attitudinal structures of respondents and which may be biased in terms of the attitudinal structures of the developers.

Having considered some of the reasons, scientific and political, for embarking on a programme of locally-based test construction in the area of inter-ethnic attitudes, we will now make a few comments at the level of practicalities.

The major source of material for such a set of tests should come, not from precarious constructs, based on a specialist test-maker's view of the world, but from ethnic groups themselves. There is an abundancy of studies which shows that test items are perceived differently by different ethnic groups. Thus, in the development of tests a number of individual as well as social determinants should be taken into consideration. This will require an understanding of the expectations, habits, norms, roles and values of ethnic groups concerned and some reference to broader kinds of influences, such as first languages and the group's social structure. The approach could proceed by setting up panels for various ethnic groups, with the emphasis on obtaining different kinds of people as representatives of their ethnic group. These panels should be broad enough to ensure wide variability of opinion and attitude within each group, yet narrow enough to be recognisable to their members as an ethnic group. It is not easy to define the limits of a language or culture, of course, but the

primary criterion, is of mutual intelligibility and reciprocity of views within the group. There should also be adequate cover of the dynamic situations which are ethnically-sensitive in the context of multi-cultural society. The functions of these panels would be: to provide data on the inter-ethnic ideological trends; to identify item sources; to help in the testing programme; to validate items for the item bank in terms of how they see inter-ethnic relations. expressing themselves in different social situations, and how they see cultural contact between their own ethnic group and the indigenous population; to have the control of the final instrument. Updating of such tests will be responsive to and in control of the different communities for whom they are designed.

The establishment of a set of such panels, to meet from time to time as a whole group, would allow those involved to determine the extent to which (a) each group is able to "define" its own perspective through its items, (b) whether its instrument can distinguish between its own ethnic group members' perspectives and those of other groups, and (c) whether the "definition" of perspectives in the testing operation does indeed promote negotiation between the ethnically-based perspectives of different groups.

Conclusion

We have based our argument in this paper on the view that tests in the area of race are notoriously defective, and on the conviction that without a dramatic change in approach to the development of these tests, the current controversy about their value and their political acceptability will continue. In the matter of local test validation, we are, of course, in agreement with many others concerned about educational testing. In the matter of local test development, we have no doubt about the sympathy of many test specialists, but are aware that the problems of local test development have always been problems of practicability. The approach we have outlined seems practicable, though it may entail new forms of organisation in the test development community, and a devolution of power over the tests from the testing establishment to those whose lives are most likely to be affected by the forms ond consequences of testing.

Like Gumbert and Spring (1974), we recognise the trends in testing over the last fifty years; there have been substantial changes in the way the use and control of testing are viewed by academic and user

communities. They write:

> "The early commitment to efficiency and expertise was being partially replaced by commitment to popular democratic control..... The general cultural revoluation that started to take place in the middle of the century appeared to be directed at the centralised and manipulative role that major institutions had assumed in society."

These changes towards democratisation in testing are, as we have argued, under threat as central government takes an increasing interest in the potential of large-scale testing. In a society, in which "reasonable racism" is respectable, we have reason to question the wisdom of a policy which may reinstate discriminatory testing practices. The controversy over testing in the area of race has identified shortcomings in the use of tests; we do not believe that they imply that testing must be abandoned. Through such alternatives as the one we have proposed here for the case of tests of inter-ethnic attitudes, we believe that future developments in testing and test use may be both epistemologically and politically justified.

Anastasi, A. Psychological Testing. New York: MacMillan, 1959

Biesheuval, S. Psychological tests and their application to non-European peoples. In G.B. Jeffrey (ed.) The Yearbook of Education. London: Evans, 1949.

Churchman, C.W. (1971) Why Measure? in B.J. Franklin and M.W. Osborne (eds.) Research Methods and Insights. Wadsworth.

Cronbach, L.J. Beyond the two disciplines of scientific psychology American Psychologist, 1975, 30, 116-127.

Cronbach, L.J. Test validation. In R.L. Thorndike (ed.) Educational Measurement, 2nd edition. Washington D.C.: American Council on Education, 1971

Downing J. and Schlesinger, P. "Racists at the Ministry", Guardian, Nov. 23, 1976.

Friedenberg E.Z. Social consequences of Educational Measurement. In P.H. Du Bois (ed.) Proceedings of the 1969 Conference on Testing Problems: Towards a Theory of Achievement Measurement. Princeton, N.J: Educational Testing Service, 1969.

Gumbert E.B. and Spring J.H. "The Superschool and the Superstate: American Education in the Twentieth Century, 1918-1970" John Wiley & Sons, 1974.

Hamilton, D.F. Some contrasting assumptions about case study research and Survey Analysis. In H. Simons (ed.) Towards a Science of the Singular, in press.

Hartman, P. & Husband, C. (1972) A British Scale for measuring white attitudes to coloured people. Race, XIV, 2.

Jensen, A.R. How much can we boost IQ and scholastic achievement? Harvard Educational Review, 1969, 39.

Karier,C. Ideology and evaluation: In quest of meritocracy. Paper presented to the Wisconsin Conference on Education and Evaluation. School of Education, University of Wisconsin, Madison, Wisconsin. April 26-27, 1973.

Messick, S. & Anderson, S. Educational testing, individual development, and social responsibility. In R.W. Tyler and R.M. Wolf (eds.) Crucial Issues in Testing. Berkeley, Calif: McCutchan, 1974.

Moore R. Racism and Black Resistance in Britain. Pluto Press, London: 1975.

Platt, W. Policy-making and international studies in educational evaluation. In A.C. Purves and D.U. Levine (eds.) Educational Policy and International Assessment: Implications of the IEA Surveys of Achievement. Berkeley, Calif. McCutchan, 1975

Schwartz, P. The AID/AIR test development project. Inter-African Labour Institure Bulletin, 1962, 9, 70-77

Thorndike, E.L. The psychology of the half-educated man, Harper's, April, 1920, p.670.

Warr, P.B. et al (1967) A British Ethnocentrism Scale. British Journal of Social and Clinical Psychology. 6, pp.267-277.