Oxford Review of Education, Vol. 19, No. 2, 1993

Assessing Group Differences

HARVEY GOLDSTEIN

INTRODUCTION

This paper is about the legitimacy of certain kinds of quantitative evidence, specifically differences in educational performance between groups, especially those defined by ethnicity, gender and class. The thrust is methodological, and because the quantitative evidence ultimately is dependent on particular mathematical and statistical assumptions, something needs to be said about these.

One of the useful things about mathematical and statistical models of educational realities is that, so long as one states the assumptions clearly and follows the rules correctly, one can obtain conclusions which are, in their own terms, beyond reproach. The awkward thing about these models is the snares they set for the casual user; the person who needs the conclusions, and perhaps also supplies the data, but is untrained in questioning the assumptions.

What makes things more difficult is that, in trying to communicate with the casual user, the modeller is obliged to speak his or her language—to use familiar terms in an attempt to capture the essence of the model. It is hardly surprising that such an enterprise is fraught with difficulties, even when the attempt is genuinely one of honest communication rather than compliance with custom or even subtle indoctrination. An example familiar to many concerned with testing is the use of the term ‘specific objectivity’ by exponents of the so-called ‘Rasch’ model. The use of this term leaves many casual users with the erroneous impression that it implies a sound and empirically verifiable justification for whatever conclusion are being drawn (Goldstein, 1979).

More pertinent to the concerns of this chapter, terms such as ‘test bias’ have been used by modellers to refer to group differences which have nothing necessarily to do with the common understanding of bias as distortion [1]. Some practitioners (see for example Shepard *et al.*, 1981) have attempted to inject more precision and acceptability into this term by defining test bias thus: “A test (or item) is biased if, ‘two individuals *with equal ability* but from different groups do not have the same probability of success’ on the test or item” (my italics).

If anything, such a definition clouds the issue even further since it falls back upon another term ‘ability’ which is undefined and indeed can only be defined in terms of other tests (or items) which do not exhibit ‘bias’, and the resulting circularity is fairly clear.

Just as the popularisation of cosmology seems to have been associated with an expansion of metaphors as well as the universes they describe, so have statistical modellers in education been coining a host of evocative terms in the area of assessment. We have acquired ‘authentic assessment’, ‘context-free assessment’ and other fine-sounding descriptions as well as the somewhat more pejorative-sounding terms such as 'bias'.

A useful service would be for a committee on assessment standards to set out guidelines on what terms are allowed. It would be useful if common language terms generally were banned unless they could be shown as unlikely to cause confusion.

I shall first review some of the empirical evidence on group differences.

THE EMPIRICAL EVIDENCE

The empirical evidence for gender, ethnic and class differences is both fairly extensive and also rather fragmentary. While the broad patterns of differences are present in most studies, the more subtle 'interactive' effects are less clear. Thus, in the UK, the differences between children from different occupational groups are well documented. Likewise, there is extensive evidence from the Assessment of Performance Unit (APU) surveys about gender differences in educational achievement. The evidence about ethnic differences is less clear cut, partly it seems because of the fluctuating composition of ethnic groups and definitional issues. Briefly, the picture is as follows.

Social Class

The 1946 and 1958 cohort studies provide perhaps the most extensive evidence on educational performance of children from families classified by the occupational group of the male head of household (Douglas *et al.*, 1968; Fogelman *et al.*, 1978). For reading and mathematics they show a gradient in performance from that of children in RG Social Class I to that of children in RG Social Class V. They also show a widening gap in performance as children get older. These studies also show differences in performance related to parental education, home amenities, and other social factors such as crowding. When comparisons are made between children with a socially 'adverse' combination of characteristics and those at the other extreme, quite large differences in performance emerge, equivalent to two or more years of educational progress.

Gender

The APU surveys on science, language and mathematics (1982a, 1982b, 1986) as well as other sources broadly agree. In general the pattern seems to be one where boys appear to make more progress than girls. Thus, in reading comprehension, an initial advantage in favour of girls in early primary school becomes a small advantage to boys towards the end of secondary school. Again, girls appear to do better on non-verbal ability tests than boys at the end of primary school, but this becomes reversed by the end of compulsory secondary schooling. There has been research which attempts to understand the reasons for gender differences, examining factors such as teacher expectations, examination entry policies, etc. A general review of explanations for gender differences and a discussion of differences in public examination performance in written and coursework components can be found in Stobart *et al.* (1992a, 1992b).

There are some interesting differences for more narrowly defined curriculum areas. Thus, at the end of primary schooling, boys are more confident in measurement and practical tasks, whereas girls perform better on computational tasks. In science

(Murphy, 1989) girls seem to be more sensitive to the need to take account of the real life context of science than boys who are more concerned with formal structures.

Ethnic Group

Plewis (1988, 1991) has reviewed much of the evidence of performance differences between black (Afro-Caribbean) and white children in the UK. There is less evidence on differences between other ethnic groups.

At the end of secondary schooling, most of the evidence is in the form of public examination results. Recent work on the results of children in inner London schools has compared the progress, during secondary schooling, of different ethnic groups (Nuttall *et al.*, 1989). The black children and the white children make very similar progress, although the black children actually tend to do worse in the final examinations, whereas the Asian groups, especially the more recently arrived groups such as those from Bangladesh, make much greater progress. There is little evidence from this work of differential progress for black girls and white girls as compared to black boys and white boys. This does appear to occur, however, during the primary, especially infant, school years where the difference in progress between black girls and black boys appears to be greater than that between white girls and white boys, for reading and mathematics.

I will argue that the existing empirical evidence has limited value. I will suggest that observed group differences are a consequence of the kind of assessment instrument used, for example, whether it is multiple choice which seems to favour boys, and that the construction of the assessment itself is influenced by expectations of what the differences are. If these arguments are accepted then some important reappraisal of educational assessment is needed.

THE GOLDEN RULE CASE

To illustrate the issues I will describe briefly a dispute which arose between the Golden Rule Insurance Company of Illinois and Educational Testing Service (ETS) (Anrig, 1988; Goldstein, 1989).

The Golden Rule Insurance Company managed to persuade ETS to adopt a policy of item selection for its entry tests which minimised black–white differences. It worked by ETS choosing a pool of items, all of which satisfied standard criteria for test inclusion. From this pool the final selection was made by choosing those items which produced the smallest (on average) differences between blacks and whites. After some years of this ETS decided that the whole thing had been a mistake and that they wished to call off the deal. The ostensible reasons were to do with the technical feasibility of administering the procedure. This created something of a stir, one useful consequence being that the issue received some exposure, at least among the testing profession.

There were certainly some technical difficulties in the procedures, but the predominant reaction from the psychometricians was that technical criteria alone, such as high reliability or correlational validity, should determine test content: one way or another, there should be a technical solution to the problem of ethnic differences (see, for example, Linn & Drasgow, 1987).

The key dispute here is between the established psychometric tradition of seeking technical solutions to problems of equity, and the proposition that it is also legitimate to seek a 'political' or 'social' solution. In the case of Golden Rule, the political consideration arrives once technical procedures are exhausted, but there is no reason why political or social desiderata should not be introduced at an earlier stage of the process.

ITEM ANALYSIS AND SELECTION

The standard psychometric approach to the production of a test or assessment, is first of all to devise items or questions from which a final selection will be made. In the setting of public examination assessments, and to a large extent in the new National Curriculum assessments in the UK, there is no stage when items or questions are empirically piloted. This is partly because there is a requirement for secrecy, and partly because to do so would be time consuming and expensive. In these cases, therefore, the suitability of the items or questions relies very heavily on their validity, which is discussed below.

The procedure for devising such items will vary, and involves elements of judgement by 'experts', and modification of existing items. It is at this stage that subjective elements of choice will enter, usually in an uncontrolled manner. Where items can be piloted prior to a final selection, the test constructor will rely on mainly statistical procedures to eliminate 'unsuitable' items. The text books on test construction typically pay little attention to the problems of initial choice, preferring to devote most attention to subsequent techniques of 'item analysis'.

While the techniques of test construction have changed over the last 70 years from a heuristic examination of individual responses to items to mathematically sophisticated modelling (Goldstein & Wood, 1989), the underlying intentions have remained remarkably constant. The following description of test construction procedures is limited to tests where piloting and revision are possible. Thus, for example, it generally will exclude the construction of examination papers for courses or syllabuses. In essence the stages are as follows.

Item Analysis

Following an initial selection of an item 'pool', and in relation to equity considerations, items will be screened for obvious biases, looking, for example, at gender or racial stereotypes in language or pictures. A prototype test or tests will be piloted on a sample of individuals, ideally drawn from the same population as the final 'target' one.

Following this, the patterns of responses to the items will be examined in detail. This stage has two principal aims. The first is to eliminate items which contain little information, for example, those which everyone gets correct or fails. The second aim is to identify 'discrepant' items prior to eliminating or modifying them. It is this second aim which is suspect.

In the absence of any external criterion against which to evaluate the test items (which I shall return to below), only the relationships among the item responses themselves are usable. To judge whether any single item is a candidate for exclusion or modification, the standard assumption is that all the items should in fact be measuring the 'same underlying attribute'. This can be defined in a more precise manner statistically using what is in effect a special case of a general factor analysis model [2] with a single underlying factor. It implies that, apart from chance fluctuations, the response on any one item can be fully predicted from the responses on the remainder.

This leads, for example, to an examination of the correlation between each item and the total score derived from summing the item responses. Items with high correlations are said to have high 'discriminations' and those items with significantly low values are then candidates for further study.

The difficulty with this procedure is that its results are sensitive to the initial choice of items. For the sake of argument, suppose that a test of reading comprehension measures two underlying attributes. Suppose also that only a few items in the test actually reflect one of these attributes. The subsequent item analysis will then tend to assign those items low correlations simply on the grounds that they are different from the majority. Excluding them from the final test will help to ensure that test only measures a single attribute. The problem is that this may not be what is required.

In more complex cases, where there are several attributes involved, the item analysis procedure cannot be guaranteed to produce anything sensible at all. The testing textbooks, by and large, attempt to make a virtue out of this unfortunate necessity by declaring that all tests have to reflect only a single underlying attribute in order to have legitimacy. What this really means is that the set of procedures used requires the *assumption* that a test reflects a single underlying attribute in order to have any logical validity. Such a requirement, however, is a very strong restriction on any assessment instrument and not one for which there is much of a substantive educational justification (Goldstein & Blinkhorn, 1977).

This use of an over-simple statistical model to determine the content of assessments pervades the testing literature. It arises in different disguises in a number of areas which I shall now explore.

Validity

Loosely speaking, an assessment's validity is the extent to which it measures what it claims to measure. Sometimes validity is measured in terms of the correlation between a new test and an old established one—the higher the correlation the higher the validity. Sometimes a test is correlated with an external criterion which is supposed to be itself a valid measure. In addition, or instead of such correlational measures, the items are judged by those designing or using them more or less subjectively in terms of their fitness for purpose. In the case of public examinations and National Curriculum assessment in the UK, the fitness for purpose judgement is paramount.

All of these procedures suffer from similar underlying problems. In the case of the correlational measures the strong assumption has to be made that the criterion itself possesses a high validity. In fact, in the case of a new test replacing an old one, it seems difficult to justify the former on the grounds of a high correlation with the latter which presumably is felt to possess important deficiencies. It is not difficult to see how, if this kind of criterion is adopted seriously, historically determined group differences can come to be perpetuated. Thus, for example, given the well-documented evidence (Gould, 1981) for ethnic differences in early tests of ability, the ethnic differences observed in current ones may, at least in part, simply be the consequence of applying this psychometric constraint when developing new tests.

In the case of the 'face validity' judgements of test constructors and users, there has to be an assumption that they have a valid understanding of how a test item or question relates to an imperfectly articulated attribute. Yet test constructors and users are themselves conditioned in their expectations by existing evidence. If they believe, for whatever reasons, that boys really do better, for example, on spatial mathematics items, it is hardly surprising if they then tend to reject those spatial items which favour girls. Gould (1981) provides a good example of a similar mechanism operating among the late 19th-century craniometrists, meticulous scientists who nevertheless were strongly influenced by their cultural expectations when forming judgements. Again, therefore, it is easy to see how historically determined patterns can persist.

In some cases at least test constructors have been confronted with having to choose between items in a test, some of which favoured one group and some another (Goldstein, 1987). We have little systematic evidence of how decisions are taken in such cases, and the fact of such choices having been made is almost never recorded.

If the above arguments are accepted, they cast some doubt upon the validity of historical comparisons of group differences. Since the assessments used generally change over time, it may well be the case that the new assessments have built in some of the previous observed group differences in order to satisfy 'validity' requirements. Thus, for example, similar (standardised) group differences over time may reflect the intensions of the test constructors as much as any external 'reality'. This also raises more general issues of how changes over time can be interpreted, but I will not explore these here.

Bias

I have alluded to the definitional problem of bias and why the standard psychometric criterion is inadequate. The question naturally arises as to whether there is *any* sense in which the term can be used.

If an assessment is designed with items that are set in contexts familiar to one group and not another, we would, in common usage, normally think of such an instrument as biased. A defence against bias would be that the context was germane to that which was being assessed, and a legitimate debate could occur on this issue. This has been a topic for discussion in mathematics education, where there is an acceptance that problems should be presented in 'real-life' contexts. For example, if such contexts are more familiar to boys than girls, then the former will tend to do better than the latter for this reason. Yet a change may well affect the relative performance of boys and girls. The question immediately arises as to what contexts to use, and I will return to this issue in the next section.

A related set of issues is raised by the science example alluded to earlier (Murphy, 1989). Here, a problem on comparing conductivities of different materials was set in the context of using the materials in clothes to be worn when walking on hills. Whereas the boys tended to ignore the 'real-life' setting, the girls were concerned with what would happen if it rained and the clothes got wet, etc. In this case, the procedure for judging the assessment might be said to be biased against the girls (or any other group with such responses) if it failed to give credit for such observations. Murphy (1991) elaborates on the different types of solutions boys and girls bring to problem solving tasks, in particular their relative unwillingness to abandon an ostensible 'real-life' context in favour of an abstracted technical issue. The issue is to decide just what is relevant to judging a response.

In these examples, there is no clear-cut solution available for judging whether bias exists or not. There would certainly seem to be a case for using the term when there is a clear *intention* on the part of the test constructor (which might have unconscious origins) to produce particular group differences. Short of this, however, the term 'bias' can no longer sustain its common meaning, and my suggestion is that it is dropped from use. We can talk more accurately about group differences or differential performance [3].

Returning to the psychometric definition of 'bias' quoted in the first section, that: "A test (or item) is biased if, 'two individuals *with equal ability* but from different groups do not have the same probability of success' on the test or item" (my italics), we can see not only that it has an inherent circularity, but also that its use can result in a subtle obfuscation of important issues. The use of the term 'ability', or indeed any other term such as 'attainment', is fraught with problems. Since all ability measures incorporate group differences it is difficult to see how any other assessment can be judged against them. Moreover, different ability tests will incorporate different sized group differences.

The appropriate role for statistical models is to summarise existing relationships, for which substantive explanations can be sought. This applies especially to the process of test construction, where they can summarise patterns in order to inform, but not determine, the process of construction.

DESIGNING DIFFERENCES

I have argued that where group differences are shown by particular assessments, these cannot be taken at face value but should be seen as characteristics of the assessments themselves, or rather of the interaction between the assessment and the groups. How then should assessments be constructed? It is important, first to separate out what might be termed a research activity, involving assessment, where the aim is to investigate why certain group differences exist, and especially in the area of gender differences there has been work in this area (Foxman *et al.*, 1990). On the other hand, where assessments are used, for example, for selection or certification, the issue is more pressing. To illustrate the problems, consider the issue of selecting children for secondary education.

The Equal Opportunities Commission (EOC, 1982) has stated that 'any allocation made (to schools or streams) should be solely on the grounds of *ability*' and that separate sex norms should not be used (my italics). Several Local Education Authorities have had to abide by the letter of this guideline (Goldstein, 1987), but the real issue is more complex.

The situation in several LEAs operating 11+ selection has been that the standard, usually non-verbal or verbal ability, tests produce higher mean scores for girls than boys. Thus, an LEA which wished to select equal proportions of boys and girls for grammar school education would be obliged to have separate cut-off points on a common test with that for girls being higher than that for boys. One consequence is that there will be some girls who fail to get into grammar schools even though they have scored higher than some boys who are selected. It is for this reason that the EOC ruled in favour of common norms and cut-off points.

The problem, however, is that an LEA which wished to subvert the EOC's intention could ask a test constructor to design a new test which, as far as possible, equalised the score distribution for girls and boys. The use of such a test would presumably not transgress EOC guidelines so long as a common cut-off was used. Yet it could achieve the same end result as having separate norms. Thus, simple attempts to legislate in this area can be effective only if they address question about the educational outcome desired, including the nature of the assessments to be used.

It would be perfectly possible to require all selection procedures to select (on average) equal proportions of boys and girls. This might be justified on social, educational, political or administrative grounds in order for it to become generally acceptable. The point is that test construction technicalities are of secondary importance compared to the choice of desired outcome. Furthermore, when explicit discussion of outcomes is absent there is always a set of *implicit* decisions being made which will determine the outcomes. For the reasons I have given, these are not necessarily the same as those which might follow from a rational debate about outcomes.

It is sometimes suggested that by adopting so-called 'criterion referenced' test design the problems associated with group differences can be overcome. This shifts the emphasis and responsibility further away from *post hoc* decisions about whether bias exists. Yet it fails to address the issue of differential expectations for different groups. Murphy (1991) suggests that boys and girls should be given different assessments, tailored to their different response styles. She implies that because such assessments would be criterion referenced, they would therefore be comparable. The trouble with this is that we cannot have context-independent assessment, whether norm or criterion referenced. If separate assessments for groups are applied there could be no proper statistical basis for equating responses, and the use of judges to compare performances would possess all the difficulties already discussed.

In the UK and several other nations it may be possible to have useful discussions about the acceptability of designing assessments with specifically determined gender differences (or rather lack of them), because there already exists experience of legislation on equal gender opportunities. The issue when applied to ethnic minorities, however, would seem to raise greater problems. The Golden Rule case is perhaps the nearest public debate which has occurred on this, and that clearly raised uncomfortable issues.

The issue seems easier to deal with in the standard educational context of an assessment which is designed to test learning in response to following a specific curriculum. This is the case, for example, with public examinations in the UK, but not with many 11+ selection tests. Even in the former case, however, there is evidence for group differences not directly related to subject-matter. For example (Murphy, 1982), girls tend to do relatively worse in examinations when multiple choice format questions are used.

IMPLICATIONS

The thrust of my argument has been that the business of constructing assessment instruments is a complex one involving social and political assumptions as well as technical manipulations. I would also suggest that because of the difficult issues this raises, it has not been easy to provoke a public debate. Added to this is the power and influence of the 'testing industry', especially in the USA and its scientific dependencies. This industry thrives, at least partly, on the need to invent and maintain sophisticated procedures for producing and modifying tests and assessments which underpin a large part of education and training. Indeed, the typical response, as in the Golden Rule case, when faced with a 'political' challenge, is to attempt to devise ever more sophisticated technical devices to deal with it.

Procedures such as that used in the Golden Rule case could usefully be incorporated into standard assessment construction techniques, including the single-occasion examinations. There seems to be no reason why the principle of trying to abolish (or otherwise constrain) group differences should be limited to gender and ethnic groups; individuals can be classified in any number of ways. Of course, practical considerations will be important, as will current political priorities. One merit of having a debate about such proposals is that it would stimulate an appraisal of existing assessments and their characteristics.

There also is another area where useful research could be carried out. Because the outcomes of assessment affect self image and the views of others such as teachers about attainment, we can think of setting up studies which deliberately modify assessments to enhance the performance of different groups. Thus, for example, mathematics assessments which included items tending to favour girls could be contrasted with those which did not. In an experimental situation, the effect of using these on student progress could be studied.

The series of rapid alterations in curriculum and assessment programmes imposed upon the British education systems in the late 1980s and early 1990s has resulted in successive large-scale assessments with different formats, contents and aims. These could also provide useful data for studying factors associated with gender differences.

Such kinds of research into group differences should make it possible to achieve a greater understanding of why groups differ, and the results of deliberately tailoring assessments to achieve particular outcomes. Such knowledge may not tell us precisely what to do, but it ought to make the consequences of any assessment decision more predictable. It may also cause us to change our minds about priorities. After all, perhaps we should accept that Murphy's girls were right to be concerned with the real-life consequences of their designs, and that less priority should be given to formal as opposed to contextual understanding. Such a choice might or might not, in the long run, advantage girls, but at least this and other similar investigations will have forced us to examine what we are assessing from a different cultural, social or political standpoint and to allow ourselves the opportunity to benefit from such experiences.

ACKNOWLEDGEMENTS

I am grateful to Desmond Nuttall, Alison Wolf and Geoff Woodhouse for valuable comments on a draft of this paper.

NOTES

[1] To make matters worse, the term is also used by testers in its everyday sense.

[2] A common procedure is to fit what is known as an Item Response Model (Lord, 1980) which is a one-dimensional factor analysis model with simple (0,1) responses corresponding to failure or success on each item. The 'Rasch' model is a simple version of this.

[3] In recognition of the problems with the term 'bias' the term 'differential item functioning' (DIF) has been coined to describe group differences for the responses to an item which differ from the common pattern of responses.

REFERENCES

ANRIG, G.R. (1988) ETS replies to Golden Rule on 'Golden Rule', *Educational Measurements: Issues and Practice*, 7, pp. 20–21.

DOUGLAS, J.W.B., ROSS, J.M. & SIMPSON, H.R. (1968) *All Our Future* (London, Peter Davies).

EQUAL OPPORTUNITIES COMMISSION (1982) *Do You Provide Equal Opportunities?* (Manchester, EOC).

FOGELMAN, K.R., GOLDSTEIN, H., ESSEN, J. & GHODSIAN, M. (1978) Patterns of attainment, *Educational Studies*, 4, pp. 121–130.

FOXMAN, D., RUDDOCK, G. & MCCALLUM, I. (1990) *APU Mathematics Monitoring 1984–88 (Phase 2)* (London, Schools Examination and Assessment Council).

GOLDSTEIN, H. (1979) Consequences of using the Rasch model for educational assessment, *British Educational Research Journal*, 5, pp. 211–220.

GOLDSTEIN, H. (1986) Gender bias and test norms in educational selection, *Research Intelligence: BERA Newsletter*, May 1986, pp. 2–4.

GOLDSTEIN, H. (1989) *Equity in Testing after Golden Rule* (Institute of Education, ERIC Clearing House).

GOLDSTEIN, H. & BLINKHORN, S. (1977) Doubts about item banking, *Bulletin of the British Psychological Society*, 30, pp. 309–311.

GOLDSTEIN, H. & WOOD, R. (1989) Five decades of item response modelling, *British Journal of Mathematical and Statistical Psychology*, 42, pp. 139–167.

GOULD, S.J. (1981) *The Mismeasure of Man* (New York, W. W. Norton).

LINN, R.L. & DRASGOW, F. (1987) Implications of the Golden Rule settlement for test construction, *Educational Measurements: Issues and Practice*, 6, pp. 13–17.

LORD, F.M. (1980) *Applications of Item Response Theory to Practical Testing Problems* (Hillsdale, New Jersey, Erlbaum).

MURPHY, P. (1989) Assessment and gender, *National Union of Teachers Education Review*, 3, pp. 37–41.

MURPHY, P. (1991) Assessment and gender, *Cambridge Journal of Education*, 21, pp. 203–214.

MURPHY, R. (1982) Sex differences in objective test performance, *British Journal of Educational Psychology*, 52, pp. 213–219.

NUTTALL, D.L., GOLDSTEIN, H., PROSSER, R. & RASBASH, J. (1989) Differential school effectiveness, *International Journal of Educational Research*, 13, pp. 769–776.

PLEWIS, I. (1988) Assessing and understanding the educational progress of children from different ethnic groups, *Journal of the Royal Statistical Society*, A, 151, pp. 316–326.

PLEWIS, I. (1991) Pupils' progress in reading and mathematics during primary school: associations with ethnic group and sex, *Educational Research*, 33, pp. 133–140.

SHEPARD, L., CAMILLI, G. & AVERILL, M. (1981) Comparison of procedures for detecting test item bias with both internal and external ability criteria, *Journal of Educational Statistics*, 6, pp. 317–375.

STOBART, G., ELWOOD, J. & QUINLAN, M. (1992a) Gender bias in examinations: how equal are the opportunities?, *British Educational Research Journal*, 18, pp. 261–276.

STOBART, G., ELWOOD, J., HAYDEN, M., WHITE, J. & MASON, K. (1992b) *Differential Performance in Examinations at 16+: English and mathematics* (London, University of London Examinations and Assessment Council).

Correspondence: Professor Harvey Goldstein, Department of Mathematics, Statistics and Computing, Institute of Education, University of London, 20 Bedford Way, London WC1H 0AL, UK.