50

to include in these the Health Centres and other community centres before we can have a real community psychologist service and full collaboration with general practitioners in the manner I have outlined.

Summary

There were 216 replies to a questionnaire planned to identify work done by clinical psychologists with general practitioners. The response indicates great interest in community psychology and desire for training in this area but also great need for further expansion both in terms of financial provision and variety of approaches to the work.

Acknowledgements

This work was made possible by a grant from the Faculty of Science and Engineering.

Dr R. Cochrane advised on the computer analysis and Professor P. L. Broadhurst, Dr M. P. Feldman and Dr J. S. Marzillier commented constructively on the manuscript.

To these and the very many who encouraged and showed interest in and discussed the survey I am indeed grateful, but especially to 216 clinical psychologists who filled in the questionnaire.

References

Bradshaw, P. W., Ley, P., Kincey, J. A. & Bradshaw, J. (1975). Recall of medical advice: Comprehensibility and specificity. *Br. J. soc. clin. Psychol.* **14**, 55—62.

Broadhurst, A. (1977). Clinical psychology in the community: A survey of general practice contacts. Submitted for publication.

Cowen, E. L. & Zax, M. (1968). Early detection and prevention of emotional disorder: Conceptualizations and programming. In J. W. Carter, Jr. (ed.), *Research Contributions from Psychology to Community Mental Health*, pp. 46—59. New York: Behavioral Publications.

Haywood, H. C. (1976). The ethics of doing research . . . and of not doing it. *Am. J. ment. Defic.* **81**, 311—317.

Kelly, J. G. (1968). Toward an ecological conception of preventive interventions. In J. W. Carter, Jr. (ed.), *Research Contributions from Psychology to Community Mental Health*, pp. 76—99. New York: Behavioral Publications.

Kincey, J., Bradshaw, P. & Ley, P. (1975). Patients' satisfaction and reported acceptance of advice in general practice. *J. R. Coll. Gen. Practit.* **25**, 558—566.

Langer, E., Janis, I. L. & Wolfer, J. A. (1975). Reduction of psychological stress in surgical patients. *J. exp. soc. Psychol.* **11**, 155—165.

Ley, P. (1976). Improving doctor-patient communication in general practice. *J. R. Coll. Gen. Practit.* **26**, 171.

Ley, P. & Spelman, M. S. (1967). *Communicating with the Patient*. Worcester: Staples.

Moore, M. F., Barber, J. H., Robinson, E. T. & Taylor T. R. (1973). First-contact decisions in genera practice: A comparison between a nurse and three general practitioners. *The Lancet* April, 817—819.

Requests for reprints should be addressed to Dr A. Broadhurst, Department of Psychology, University of Birmingham, Birmingham B15 2TT.

Bull. Br. psychol. Soc. (1977), **30**, 309—311 *Printed in Great Britain*

Monitoring educational standards — An inappropriate model

Harvey Goldstein and Steve Blinkhorn

A major element in the DES proposals for the 'raising of educational standards' is a system for the regular assessment of educational attainments. The Assessment of Performance Unit (APU) within the Department has been given the task of carrying out such monitoring procedures in schools. Naturally, it has considered the various problems of devising adequate measuring instruments and agreeing upon measurement of common objectives, and a large amount of public discussion of these issues has taken place. In the context of the current 'Great Debate' there has even been some discussion of whether we should be trying to carry out such a programme at all.

The present notes are not directly concerned with these issues, although what we have to say does bear on them. We would like to make it clear that we are not objecting in principle to routine monitoring using standardized assessment techniques. Rather we wish to express certain doubts about the particular statistical methodology which it is proposed to use, and to discuss certain pedagogical implications. We will present what we have to say in a non-technical fashion, relegating the one technical detail to a short Appendix.

Item banks

As soon as one begins to think about assessing attainment one meets the problem of how to devise an instrument which provides fair comparisons between children subject to the styles, methods and contents of different teaching regimes. It is well known that educational objectives vary between teachers and with time. In particular, previous attempts to compare standards over time have been criticized for failing to take account of changing curriculum objectives which become reflected in children's performance. Recognizing this problem the APU has suggested 'item banking' as one solution (Kay, 1976).

The idea of such a bank is that one simply selects a sample of items from the bank with known difficulty values, and estimates an individuals' attainment or ability by noting the responses to these items. The assumption here is that a test item has a single 'difficulty' value and that different items which are appropriate for assessing different educational objectives are strictly comparable along this single dimension. With existing tests such a simple situation is not true in general, although in certain circumscribed areas it may be approached. The novel claim for item banks is that the methodology now exists to bring this about.

According to Willmott & Fowles (1974): 'The Danish mathematician George Rasch has proposed a model whereby objective measures of attainment can be achieved (Rasch, 1960). The word 'objective' is used to mean that the measurement can be developed and reported without reference to a particular set of items or a particular group of people. . . . It would no longer be necessary to obtain representative samples for pretesting purposes if objectivity were achieved.' These are strong words indeed, and if true the Rasch model proposed might well be seen as the philosopher's stone of the psychometricians! To evaluate this claim, however, we need to describe this particular Rasch model and we do so here in a non-technical manner. The Appendix presents a mathematical formulation of the model.

The Rasch model

Suppose we present a set of items to a group of individuals and score each individual's response to each item as a 'pass' or a 'fail'. We can record these in a table such as the following.

	Items			
Individuals	1	2	3	. . . n
1				
2				
⋮				
m				

Each cell of this table will contain a 1 (pass) or 0 (fail). Now consider the probability that a given individual passes a given item. This will, in general, depend on the ability or attainment of the individual and also the 'difficulty' of the item. The Rasch model *assumes* that any given item has the same relative difficulty for each individual so that in order to calculate the probability of a pass we need to know only the individual's ability and the difficulty value of the item. The possibility of an 'interaction' between individual ability and item difficulty is ruled out, for example that the order of difficulty may be different for high and for low ability children or for children subject to different teaching methods. Further, dependencies between items are assumed not to exist, an assumption which is particularly questionable when items are selected from a bank (see Appendix). If the model is accepted then it represents a simplification in the sense that $m \times n$ individual responses can now be expressed in terms of $m + n$ individual abilities and item difficulties, these quantities being calculated by a sophisticated averaging process over the cells of the table.

If, for the moment, we assume the model to be true for a group of individuals, does it in fact possess the properties claimed for it? Clearly not. The fact that it is a good representation of a particular set of items for a particular group says little in itself about whether it will do so for other items and other groups of individuals. To establish this is a matter for empirical investigation rather than axiomatic definition. For example, the particular order of items for a group of children taught 'new' maths might be quite different than that for those taught 'traditional' maths, even though the items themselves could be regarded as validly measuring mathematical ability. Likewise, the order of difficulty for a set of items may change over time. In the quotation given above there seems to be a certain confusion between *objectivity* and *simplicity*. The Rasch model represents a particular samplification of the observed responses, but this does not entitle us to claim objectivity, even in the restricted sense in which this word is used by Willmott & Fowles (1974).

It is by no means clear that the Rasch model does describe real data very well. Willmott & Fowles (1974) admit that when testing the model some items do not fit the model. *These are omitted from the set of items.* As they say, 'The criterion is that items should fit the model, and not that the model should fit the items.' (!) Thus we have a measuring instrument composed of items which conform to a particular mathematical model, and moreover as has been indicated, one which presupposes a highly simplified view of cognitive functioning. The adoption of such a model for the construction of item banks to be used in routine assessment could pose serious dangers for the educational system. Only certain types of items will be included in the bank, and teaching methods may be changed in order to maximize pupils' chances of success on these items.

For example, it seems plausible to suppose that a child taught new maths might have less difficulty doing binary arithmetic items than a child taught by more 'traditional' methods. Conversely, the former child might be at a disadvantage with items involving rote multiplication. If this were the case then both kinds of items would tend to be excluded from the bank since they would not have comparable difficulty values for children taught by each method.

To some extent, of course, similar consequences can flow from any test, however constructed, if used for regular routine assessment. The particular danger with an item banking procedure lies in its *claims* to be 'objective', 'sample independent' and a solution to the problems of comparability over time. This is compounded by the sophistication of the mathematical model used, whose details will be inaccessible to the vast majority of those involved in discussing educational policy, and who are therefore unable properly to evaluate these claims.

Finally, we should reiterate that we are not objecting in principle either to the use of mathematical models in psychological research, or to the principle of standardized assessment. What we wish to emphasize strongly, however, is that these models are not yet in a fit state to be used other than as interesting research tools. We believe that their claim to 'objectivity' is unfounded and that the requirement that data conform to a particular mathematical model creates effects which are largely unknown. We believe that the present attempts to apply them in a *routine* manner to monitor educational standards should be viewed with caution. If we are to try to measure standards we need to think out very carefully the objectives and implications of any instruments we use.

Acknowledgements

We would like to thank Mr M. Couzens, Professor C. Hindley, Dr B. Kay, Professor W. Wall, Dr A. Willmott and Dr R. Wood for their helpful comments.

References

Anderson, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika* **42**, 69—81.

Kay, B. (1976) Justified impatience. *Times Educational Supplement* 1.10.76.

Rasch, G. (1960). Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danmarks Peadagogiske Institut.

Willmott, A. S. & Fowles, D. E. (1974). The Objective Interpretation of Test Performance. Slough: NFER.

Appendix

The proposed model can be written

$$\log \frac{P_{ij}}{1\text{-}P_{ij}} = \beta_i + \gamma_j ,$$

where β_i refers to individuals and γ_j to items for the two-way cross-classification described in the text. Typically the β_i are positive and the γ_j negative. P_{ij} is the probability of a pass for individual i on item j. Thus we have an additive model on the logit scale and estimation procedures are fairly straightforward. For a further discussion of this model and the extension to multiresponse items see Anderson (1977).

Local independence is assumed, i.e. for a given individual the response to any item is independent of the response to any other. That is to say, no account is taken of possible facilitating or inhibiting effects of responses to prior items on responses to subsequent items. Such an assumption is common in test theory, but assumes particular importance when a test is derived solely on the basis of the difficulty values of items from a bank. This procedure does not involve the traditional process of validating a complete test during which the local independence assumption can be approximately satisfied.

Requests for reprints should be addressed to **Harvey Goldstein**, University of London Institute of Education, Bedford Way, London WC1.

Steve Blinkhorn is at The Department of Psychology, Hatfield Polytechnic.

Careers in Psychology

What exactly is a psychologist?
What do they do?
How do I become a psychologist?
How long will it take?

The British Psychological Society has prepared this booklet to give information on these and many other questions which will be asked by those planning entry to a university or other institution of higher education who may be interested in taking psychology as a main subject in an honours degree. The booklet should also provide much useful information for those involved in careers and occupational guidance.

Price: 50p

Available from:
The British Psychological Society
St Andrews House
48 Princess Road East
Leicester LE1 7DR