1978
NCMt

THE NATIONAL MONITORING OF ACADEMIC STANDARDS

by

Bruce Choppin

National Foundation for Educational Research in England and Wales

paper read at a meeting of the
National Council on Measurement in Education

Toronto

March 1978

The National Monitoring of Academic Standards

by

Bruce Choppin

National Foundation for Educational Research in England and Wales

Introduction

It seems appropriate to report at this time on progress of a British program for the National Monitoring of Academic Standards for two reasons. First, the measurement problems involved in such an exercise are complex and even our first thoughts at finding solutions to them may be of interest. Second, to a very considerable extent the strategies we are adopting in Britain for tackling these problems and others derive from the experience of the National Assessment of Educational Progress (NAEP) in the United States over the last decade.

To begin with it is perhaps best to look at just why Britain decided two years ago to begin this monitoring program. The main cause is undoubtedly that school standards have become an extremely emotive issue following fights over the structure of our educational system in the political arena. Britain's traditional selective system of secondary education, in which children belonging to different bands of ability went to different types of school, was largely replaced during the late 1960's and early 1970's by a system of comprehensive high schools, each one catering for the full ability range. This change-over is still not complete but now the vast majority of British children of secondary school age are in the comprehensive system. The change has been vigorously resisted by those on the political right, whose principal claim has always been that the comprehensive school would lead to a lowering of standards. In recent years this argument has been embellished by reports of large numbers of pupils leaving school without having mastered even minimal standards of literacy and numeracy, and by largely inaccurate claims that research evidence has shown that standards in reading and mathematics achievement

are falling. The truth is that we have no adequate data to determine whether average standards have risen or fallen over the last decade. We can say, on the basis of standardised test scores, that immediately after World War 2 and well in to the 1950's, standards were rising quite appreciably. What is more, children were tending to stay at school longer and more of them were going on to some form of post-school study. These rapid increases are no longer taking place. That much is clear, but we do not have any data which would tell us whether standards are rising or falling at the moment. All we do know is that the changes are small. This of course is not satisfactory from the point of view of politicians or public opinion. They are less interested in the speed of change than in its direction.

The government therefore decided to establish within its Department of Education and Science an Assessment of Performance Unit which would mount studies to establish a bank of reliable information about the levels of achievement, and which would serve as a basis for the monitoring of future changes. Substantial parts of this work have been subcontracted to the National Foundation where I am employed.

The scope of the monitoring work will be wide. It is eventually intended that most of the curriculum will be included. So far detailed plans for the testing of mathematics, English and Science have been developed covering both primary and secondary schools. The field work will begin this May with the testing of a sample of primary school children in mathematics.

The Aims of Monitoring

A major problem with regard to the NAEP which has still not been fully resolved concerns the purpose of the monitoring itself. In late 1963, the first list of major objectives of the national assessment included:

a) to gather data descriptive of the strengths and weaknesses of the educational system and to collect this data periodically so as to provide a census of educational progress.

b) to aid Congress and the general public in arriving at more informed decisions on policy issues.

c) to support and assist researchers working on various teaching and learning problems.

You may recall that the original design of the National Assessment was drawn up by Ralph Tyler and John Tukey, but the earliest statement of objectives was influenced by a wider group that included John Gardner (from a research foundation) and Francis Keppel from the USOE. Greenbaum, in his analysis of the NAEP points out that although Gardner and Keppel needed the second and third objectives to justify the exercise, there is no evidence that Ralph Tyler ever really subscribed to them or regarded them as realistic. Tyler believed that the National Assessment could achieve the "periodic census" objective, and that such data would be of genuine though limited use. Greenbaum in his evaluation argues cogently, and at length, that, given the agreed design of the National Assessment, only objectives in the first group were really feasible, but that it has taken a very long time

for this to be generally accepted. He puts the blame on the staff of the NAEP itself (and presumably Ralph Tyler) for raising expectations about the outcomes of the National Assessment to unrealistic levels. The widespread disappointment at the seeming impotence of the NAEP to answer "the burning educational questions of the day" is the result.

I mention this because we have a not entirely dissimilar problem in Britain. Of course we have had the opportunity to learn from the experiences of the United States, and it is noteworthy that a very thorough review both of the NAEP and of State monitoring schemes preceeded the design of the British program. Nevertheless arguments have sprung up in various places about the design of our assessment; about whether it is too ambitious or not ambitious enough; about whether the results may have bad effects or perhaps no effect at all. I sat in recently on a meeting of our Science panel where this debate was very much in evidence. Few present could see much value in a purely descriptive survey of current practices and standards. Rather than just to describe the educational scene, the enthusiasts would like to try and isolate the causes of high and low achievement so that some corrective action could be prescribed. The monitoring was clearly a unique opportunity to carry out all the research studies into the effects of various combinations of background variables that could be imagined for the next decade or so.

The inclusion or exclusion of background variables is a key feature of the ground on which the battle between the "descriptive census" and the "causal relationship" approaches is fought out. Despite the extravagant tone of some publicity handouts the National Assessment in practice adhered quite strictly to the descriptive

line. I quote from Greenbaum;

> "The Tyler-Tukey position, doubting the utility of attempts at causal analysis, prevailed virtually unchanged from the first meeting in 1963 until very recently." (p.116).

Similarly in Britain while certain subject committees, the National Press, and some educational lobbies argued the advantages and/or the dangers of the research problem/policy decision approach, the Assessment of Performance Unit has been careful to restrict the background variables collected during the first surveys to a minimum. Apart from identifying the sex and age of a child and the region of the country in which he or she attends school, it is likely that no more than three or four pieces of socio-ethno-psycho- or economic data will be collected to give some possible additional reporting categories.

It is appreciated within the APU, I believe, that the objectives and the limitations of the monitoring program need to be generally, and widely, understood if the program itself is to succeed. Realism may prove to be a better policy than extravagant and over-optimistic generalisations when countering the fears of local government officials, teacher unions and the like. As Burstall and Kay state:

> "It could be argued that NAEP is less effective than it might have been because, partly too blunt initial opposition, its political objectives were not very precisely stated, or perhaps generally agreed among its planners." (p.3).

Psychometric Issues

On the measurement issues as well the British have learnt from the experience of National Assessment. The American program has embraced matrix sampling and has made it work. Each student drawn in the sample of the study is asked to respond

to only a few test items sampled from a larger pool. On the other hand the interpretive framework employed has been less successful. The NAEP has repeatedly stated that its intention is to carry out criterion-referenced measurement, but a number of critics doubt whether this has ever been achieved. Greenbaum for instance comments

> "although the exercises are loosely related to the objectives there is no way of knowing the level of performance on the exercises that would indicate the achievement of the objective." (p. 100)

and

> "Thus while the exercises are in some general sense objective-related, they are not in any meaningful sense criterion-referenced". (p.105)

The National Assessment relies very much on a consensus as regards content validity for determining which exercises should be included. Such item analysis as was performed appears to have been of the norm-referenced type although it is perhaps worth noting that very little was reported on the National Assessment's analytical procedures in the early years.

The British Monitoring program will also employ matrix sampling, but this is intended to be part of a fully-developed item banking strategy. Test materials will be written, piloted and calibrated to form the banks (one to each subject area) with only a small proportion of the items being used for the monitoring in any one year. The reason for this is that it is accepted that a major purpose of the APU work is to detect changes in educational standards over time and also to document and describe the changes that occur in the curriculum as it is taught in the nation's schools. To avoid the norm-/criterion-referenced dilemma, latent trait theory will be invoked, and in particular we plan to scale test exercises

and pupils' performances in accordance with the model proposed by George Rasch. Reporting on the latent trait in some agreed unit, probably the wit, gives a stable basis for comparing results between groups and between different times of testing. Since individual test exercises are also calibrated on the same scale these provide a way of illustrating observed levels of achievement by reporting specific exercises that carry an appropriate level of difficulty. The reporting will thus probably avoid altogether the "mastery" syndrome and the examples given in the reports will be of exercises of a realistic level of difficulty (i.e. about 50%) for most students. I should perhaps say that these are, at the moment, no more than plans and it could be that either unforseen psychometric problems or political pressures will force us to resort to some other and more traditional form of reporting.

New problems are generated by this latent trait approach. The first concerns the dimensionality of the measurement - whether we are dealing with a uni-trait or multi-trait situation. We have firmly decided not to try and assess performance on each of a large number of separate educational objectives (the approach adopted by NAEP), but many people are unhappy with the most immediate alternative which would be to assess mathematics performance as a single-trait global entity. The team constructing the test exercises are working with a model of school mathematics which identifies thirteen separate areas. An attempt will be made to assess performance on each of these (although, within the matrix sampling scheme, each individual student will supply evidence for only three out of the thirteen). We shall also explore the extent to which the sub-trait representing performance in each area can be seen as a component of the global trait that we shall call

"Mathematics". As far as possible we wish to approach a solution in which we can report on global "mathematics" as well as on the thirteen sub-traits. Sub-trait measures however would represent deviations in performance from the overall trend and would be interpreted as showing local variations in teaching emphases etc. Whether or not we can do this satisfactorally will depend upon more detailed investigation of the first round of data which will not become available for a few months. The Rasch model is in essence a single-trait model and we will definitely need to arrive at a fairly economical description of performance in terms of traits if we are to apply the model efficiently.

In the past a great deal of energy has been expended on the debate as to whether the Rasch model fits real data (or whether real data fit the Rasch model). I had better make it clear that while "fit" will be an issue in our work it will play a quite separate role. A group of conventionally sound items, which when grouped form a test of a single trait, can be adequately represented and explained by the Rasch model. If they could not then conventional testing procedures and analysis would not work. When we construct our test exercises we use conventional item analysis procedures and Rasch item analysis, as well as careful scrutiny by teachers and others, in order to eliminate items that have obvious measurement defects. Once items pass this screening and are in the bank we do not expect that any subset of them on any particular occasion will exactly conform to the behaviour predicted by the Rasch model. It is after all, no more than a model, an approximation to the truth. The residuals in the data matrix once the pattern predicted by the model has been subtracted out provide detailed diagnostic information about the students and the items, and this information can be put to good use. We are

charged in our work for the APU to explore changes in educational performance over time, and this implies not only the measuring of increases or decreases in standards of performance, but also the monitoring of changes in what is taught. The mathematics curriculum as taught in the schools today is not the same as that taught ten years ago and will not be the same as "Mathematics" taught ten years from now. This change in the definition of "Mathematics" over time, we have called curricular drift and it is a major part of our work to identify and assess it.

In a certain sense the latent trait "Mathematics" may be thought of as being determined by the exercises that make up the mathematics item bank and thus would remain relatively constant (although a few items will be added and a few items deleted each year). More importantly though, the definition of the latent trait emerges from the interaction of students and test exercises as evidenced by the pattern of relative difficulties between exercises. As these relative difficulties change then so does the definition of the latent trait. This can be reported in terms of certain topics in the curriculum becoming easier as more emphasis is placed on their teaching while others become progressively more difficult as their importance in the classroom is reduced. We hope to be able to monitor this drift at the same time that we keep a tight check on the levels of performance being achieved. This is a bit like pulling oneself up by one's own boot-straps, and it can only be done by making a number of assumptions the feasibility of which remain to be tested in the field. It is clear, for instance, that we could not hope with this system to cope with a revolutionary change in school teaching that resulted in most of the curriculum being changed from one year to the next. Fortunately

this situation is not likely to occur. Our best estimates at the moment suggest that the overlap between the mathematics taught now in the secondary schools and that taught ten years ago is about ninety per cent, which if I may invoke an extremely crude and over simplified model suggests that about one per cent of the content may be being replaced each year. Our monitoring program runs on an annual cycle but the scale of the work (sample sizes, precision of measurement etc.) has been determined so that we would expect to be able to identify changes in standards and also curricular drift over a period of some three to five years. By 1984 we should know if our early estimates were approximately correct.

The test exercises

The final problem that I wish to outline is political and social in its origins but has serious psychometric implications for the monitoring process. We have a decentralised educational system in which central government has very little impact on the content of the curriculum. This is determined partly at a County or Borough level and partly within the individual schools. As a result, although most pupils are working towards one or other of our national examinations which are based on published and relatively fixed syllabi, there is considerable scope for the individual school especially in the earlier years of secondary education to choose which topics to teach in depth and which to de-emphasise or perhaps ignore altogether.

The question then is which test exercises should be used in a national monitoring. Is it fair to set pupils tasks which centre around topics that are unfamiliar to them? Is it fair to infer that a teacher has failed when the evidence suggests he

has not covered a particular topic, if in fact he has concentrated on other equally worthy topics which are not included in the test? It is not thought economically feasible to arrange separate test exercises for every class of children drawn in the monitoring sample so that each matches exactly what has been taught. What then can be done?

There are some pressures to restrict the testing exercise to what is seen as a common core-curriculum which will have been studied (or at least should have been studied) by everybody within the target population. This in a sense is what the NAEP had left after its extensive procedures of scrutiny both by subject matter experts and lay panels. In Britain we have been less deliberately democratic about exercise construction. Many educational practitioners harbour very real fears that the government may be looking for ways of imposing a "common core" of curriculum upon all schools*. The test exercises are being composed by panels of subject matter experts to cover a cross-section of what they (not the Government or the PTA) see as current good teaching practices in the schools. We seem therefore to be committed to a wide range testing programme which will inevitably require some pupils to attempt questions on topics which they have not been exposed. To the extent that our function is to monitor performance (that is; to measure just what students can do) this seems reasonable, but it does cast doubts upon the logic of the psychometric methods we employ. The Rasch model is, fortunately, an extremely flexible device. If we had accurate data as to which student-item interactions were "fair"

* A distinguished academic on the most senior consultative committee of the APU recently resigned in protest against moves towards this standardisation of what schools teach, a policy that in Germany between the two Wars he saw as linked to the rise of Fascism.

and which were "unfair" (in the sense that the student had not been taught the topic) then we could assess the students separately on the two sets of items. In practice this is thought to be unmanageable. Teachers are notoriously reluctant (in our country at least) to admit that they may in fact not have covered a particular topic if it is on a prescribed syllabus.

We do not see that there is a complete answer to this dilemma. The solution that is being attempted is to establish the pattern of relative item difficulties from data provided from samples of people who have been exposed to the items under consideration. Our estimates of relative difficulty and consequently the establishment and calibration of our scales of performance will be developed on a set of data resulting from "fair" interactions. Ideally, this will need to be done outside the regular monitoring programme. As the monitoring proceeds we shall use these calibrations to estimate the performance levels of individual students without regard to whether the students have or have not been exposed to all the topics. This will introduce some noise into the data which should show up as fairly random deviations from the Rasch model for certain topics and if we can identify these we shall (tentatively) report that they may not have been taught as widely as they might have been. Once again I am telling you about our plans and it is too early to say whether this will work well in practice.

Conclusion

The present paper has not got the scope to go into more detail on our testing and analysis strategy, or to consider other problems that loom in the future. We know for example that we shall be including both objective 'multiple-choice' type items

and more extended 'constructed response' type items in our assessment of language - the latter to be scored on a 0-10 or a 0-20 basis. We know how to apply the Rasch model to each group separately, but when both are combined in an item bank we expect trouble. Again we doubt that the monitoring of "aesthetics" or "personal development" when they are introduced will be amenable to the measurement techniques we are using for Mathematics and Science. In time these problems will have to be faced. Then as now we would hope to draw on the experience of other similar endeavours, and particularly the NAEP.

References

BURSTALL, C. and KAY, B. Assessment - The American Experience, Assessment of Performance Unit, DES, LONDON 1977.

CHOPPIN, B. Item Banking and the Monitoring of Achievement, Research in Progress Series No. 1. NFER Slough 1977

GREENBAUM, W. Measuring Educational Progress: A study of the National Assessment, McGraw-Hill, New York 1977.