In Mathematics versus the National Curriculum, eds Dowling, P.& Noss, R. Falmer Press, 1990 pp 67-76.

3 The Fundamental Assumptions of National Assessment

Harvey Goldstein

Introduction

Of all the innovations in the 1988 education act, arguably the most important and most influential in the long term, will be the proposed system of National Assessment. The basic framework for this was the report of the Task Group on Assessment and Testing (TGAT), published in January 1988 (DES, 1988a). Most of the Report's major assumptions were accepted and are being implemented by the present government, through the Schools Examination and Assessment Council (SEAC), a government appointed body with wide powers to regulate and develop school examinations and assessments. Academic and research institutions have obtained contracts to prepare the details, principally in the form of 'standard assessment tasks' (SATs).

The speed with which the TGAT proposals were formulated and then operationalised has left little room for discussion of the more fundamental issues around a national assessment system and, this paper argues, has led to severe problems. In the paper I shall examine the notion of criterion referenced assessment, the proposals for teacher assessment, for the reporting of results, and the issue of gender and other 'bias'.

It is worth noting that the National Curriculum has been conceived in subject terms with little serious attempt to formulate genuine cross curricular structures. This has important implications for modes of learning and to some extent also for assessment. In particular, a cross curricular perspective would make the notion of 'learning hierarchies' more difficult to sustain and would also force assessments to take more account of contextual issues. This latter issue is discussed below, but space does not allow a more detailed exploration of all the implications of a subject based curriculum.

In the following sections I will argue that the surface plausibility which adheres to these issues begins to fall away when examined closely. I will also attempt to draw some conclusions and to suggest that there are more systematically rational approaches to assessment and 'standards' than so far have been proposed.

Criterion Referenced Assessment

Criterion referenced assessment ideas feature prominently in the arguments in the TGAT report. They also appear in the reports of the curriculum working parties which were set up to formulate the detailed structures of the National Curriculum in each subject.

The idea of criterion referenced assessment became articulated in the 1960s (Popham and Husek, 1969) as an attempt to link assessment to learning objectives. In the 1980s it has seen a resurgence in the UK; in grade criteria for the new 16 year old school leaving qualification, the General Certificate of Secondary Education (GCSE), in the graded assessment movement, in some of the early work on profiling and now in the attainment targets for the national curriculum. Often crudely interpreted in terms of 'can do' statements, it is promoted as a provider of practical information about what a pupil has 'learnt and mastered' (TGAT § 94). The report's claim is that 'Norm referenced approaches conceal changes in national standards.... Only by criterion referencing can standards be monitored.' (§ 222). The report provides no indication how such 'standards' are to be derived and communicated. In any case, the difficulty (if not impossibility) of measuring changes over time has little to do with the form of the assessment or the educational philosophy behind it. The difficulty arises from the fact that an assessment used at one time will generally need to be updated periodically to reflect curriculum changes, the introduction of new technology or language, and so forth. This means that the assessment instruments or tasks change over time and no 'absolute' standard or scale is feasible (Goldstein, 1983).

Attempts to produce descriptions of 'mastery' based on criterion referenced ideas have needed to operate at a level of generality which has demanded a set of 'context free' descriptions. Thus, for example, the report on Mathematics Attainment Targets in the National Curriculum (National Curriculum Council, 1988) quotes a specimen maths attainment target as: 'Select materials and the mathematics to use for a practical task'. In reality, the information upon which any such description can be based will be limited, and to make a decontextualised

statement of achievement on such a base requires major assumptions. The single example for each target, given alongside, is hardly sufficient. In short, we have to assume that such statements can be applied in the far greater number of contexts which were not observed. What we do know is that in general this cannot be done. The work of the APU in mathematics, for example, has shown how something as simple as a change in presentation format can change performance markedly (APU 1986) and the same is true in language assessment (Thornton, 1986). There is now beginning to be some systematic research into this area, and this is exploring the ways in which learning and understanding are linked to the contexts surrounding the learner, her motivation and her perceptions of purpose (Wolf, 1987; Walkerdine, 1984).

The TGAT report, on the other hand, has no doubts. There is no recognition there that problems may exist and we are informed simply that 'the system is also required ... to play an active part in raising standards of attainment. Criterion referencing inevitably follows'. (§ 222).

Teacher Assessment and Moderation

There are two distinct kinds of assessment discussed in the TGAT report. One is a series of centrally designed 'standardized assessment tasks', both written and practical. The report spends time arguing in favour of 'innovative' and interesting tasks which can be incorporated into daily teaching. These tasks will be marked by teachers who will receive relevant training.

The other kind of assessment is that to be done by the teachers themselves on the basis of their pupils' general work and in the same 'profile component' areas covered by the centralized assessment. The report devotes much space to describing how the teachers' results are to be made compatible with each other and with the centralised assessment. The report recommends that 'teachers' ratings be moderated in such a way as to convey and to inform national standards.' (§ 62). It suggests that, if left alone, 'teachers' expectations (of what is normal) become the teachers' standards' (§ 65). The report recognizes that 'teachers' rank orders ... may vary systematically from rank orders provided by test users (§ 66)', and so the notion of teacher assessment adopted by TGAT is one where such differences are eliminated.

Where teacher assessment is a matter for discussion, negotiation

and recording between pupil, teacher and parent, then there is no requirement to convey national standards, nor indeed for teachers to agree among themselves. Furthermore, such locally based assessment is in many respects more appropriate as the basis for decisions about curriculum provision, individualized teaching schemes and so forth. It is precisely its ability to reflect local conditions which makes it valuable. It is only where comparability is paramount that the above requirement is seen to be necessary. Yet neither the first report nor the supplementary reports recognize this distinction and by implication, therefore, would seem to place lower value on those elements of teacher assessment which do not accord with the centralized assessment. A likely consequence is that teacher assessment would become restricted to just those things which can also be measured by centralized tasks. Indeed, the report itself seems to envisage this when it recommends that 'support items, procedures and training be provided to help teachers relate their own assessments to the targets and assessment criteria of the national curriculum.' (§ 116).

Since the TGAT report, the role of teacher assessment seems to have been reduced in importance, while that of the SATs has increased. It is also becoming clear to many people that assessment instruments which are designed for public reporting of results are inappropriate for 'diagnostic' assessment of learning opportunities and difficulties. If teacher assessment becomes very strongly linked to the SATs and hence of the school reporting process, it is then not very relevant for diagnosis.

Reporting School Results

National assessment is a central feature of the 1988 Education Reform Act. The proposals to use these assessments to make comparisons between schools will also become the most important part of the system.

The TGAT report proposes that, for profile components, or groupings of these, each school should report its average level (or distribution of levels) at ages 11, 14 and 16. Although it recommends against publishing the results at seven years, this has been rejected by the Secretary of State for Education who strongly recommends that each school's seven year results should be reported (DES, 1988c). The report also suggests that at the same time, a report is attached describing the socioeconomic and other characteristics of the area surrounding the school. The implication of this is that parents and other users

Table 1 Average LEA exam scores

	Rank Orderings			
	Unadjusted	Adjusted		
LEA		A	B	C
Harrow	1	1	1	4
Barnet	2	2	35	27
Coventry	59	5	7	1
Haringey	90	91	79	66

of these results will be able to make allowance for these factors when comparing the performances of schools.

Needless to say this is easier said than done. Apart from the obvious problem that there will often be a mismatch between the characteristics of a school neighbourhood and those of the children actually attending the school, it is unsurprising that the report fails to suggest precisely how the allowance is to be made. The various efforts by others in this area, notably the ILEA, have been unsuccessful. It has been known for some time that such attempts to 'adjust' or allow for influential factors solely using 'aggregated' data, are difficult if not impossible. For example, Woodhouse and Goldstein (1988) carried out such analyses for exam results aggregated to the LEA level, using data from the DES. The results, in Table 1, are based on an analysis relating average LEA exam results to socioeconomic and demographic factors.

The first column gives the rankings of the three LEAs shown using just the unadjusted exam results. Column A gives the results after adjustment for socioeconomic and demographic factors, as presented by Gray and Jesson (1987). Their statistical model was then slightly modified and column B gives the rankings thus obtained. Column C gives the rankings from a further modified model.

In terms of describing the observed data, all three models do equally well and there is no objective way of choosing between them. Yet the results for individual LEA's can vary markedly. It is, to put it mildly, somewhat optimistic of the authors of the TGAT report to suppose that very much sense could be made of its own proposals. This is especially so since the report ignores the single most important factor influencing achievement during schooling, namely the achievement of the pupils at time of entry to school. Indeed, it seems that the proposals cannot be implemented in their present form.

There is now a widespread interest in measuring school effective-

ness, using a variety of procedures, qualitative as well as quantitative, but none of them is free of problems and there is no real consensus on how, or indeed whether, it can be done satisfactorily. Most recently, some workers have advocated the use of 'multilevel' statistical models (Aitkin and Longford, 1986), but these too have their problems and there is no guarantee that they would be able to supply convincing school comparisons, even if the resources were available to utilize them. Perhaps the most extensive analysis along such lines has been that of examination results in the Inner London Education Authority (Nuttall *et al.*, 1989). This confirms that analyses which use school averages can be highly misleading. The analysis also finds that schools differ along more than one dimension so that a single 'effectiveness' measure provides an incomplete picture. Thus, for example, the average difference in exam grades between those students who are in the top 25 per cent in terms of a verbal ability test score and those in the bottom 25 per cent on verbal ability, is just under 3 A grades at O-level.[1] This difference ranges from just under 2 A grades to just under 4 A grades across schools. The average difference between students of Pakistani origin and those originating from England, Scotland, Ireland or Wales, is about one A grade, but varies from zero to 2 A grades across schools. Likewise, the gender difference varies from school to school. All these differences are adjusted for the intake achievement (on verbal reasoning) and are only moderately intercorrelated.

In existing debates on the use of so called 'performance indicators' in schools (FitzGibbon, 1990) many of the same issues of using aggregated data arise. School examination and test results may well come to form the core of such indicator systems.

The importance of taking account of 'intake' achievement to a school and using multilevel analysis, is now widely recognized as the only secure starting point for comparing schools. Even so, the best that can be expected is that 'extreme' schools will be screened out. Such schools, whether apparently markedly good or markedly bad, would be available for further investigation by inspectors and advisors in collaboration with the schools and the Local Education Authorities. It would not be possible to say anything useful about the majority of schools which do not stand out, and in no way would it be legitimate for the results of such a screening programme by themselves to be used to pass judgement.

There is now plenty of evidence, much of it from the USA, that to use assessment results to compare schools promotes wasteful and unfair competition. It leads schools to concentrate on 'playing the

assessment game', encouraging them to 'teach to the test' and to spend energy on finding ways to improve their test scores by means which are largely irrelevant to the true business of education. A spectacular example of this from the USA is the so-called 'Lake Wobegon' effect whereby every state in the USA was found to have an average test score above the national average! (Cannell, 1988).

Group Differences

Finally I would like to raise some problems of equity. I will discuss them in terms of gender, although the same general points apply to differences based on other classifications, for example ethnic groupings.

There are well known gender differences in various topic areas and according to test format. Thus, for example, girls perform relatively worse on multiple choice tests, and better on tests of 'verbal reasoning' at all ages. The scrutiny of test material for racial and sexual stereotyping is, by now, a standard procedure among test constructors.

Educational test constructors use the term 'bias' simply to refer to any item (or test) which shows differences between well defined groups. Thus, a multiple choice item might be described as biased in favour of boys if more boys obtained a correct answer, on average, than girls. Unfortunately, this procedure, which relies on the observation that some test items are more difficult for some individuals or groups of individuals, does not tell us what to do with those items. For example, should we regard the higher performance of girls on 'verbal' items as an indication of 'bias' or a 'real' reflection of girls' superiority? Should we eliminate multiple choice items on the grounds that they are biased in favour of boys?

Ultimately, the answers to such questions are political and ideological rather than technical. In general, by judicious selection, we can choose tests which on average will favour boys, or girls, or neither. We can also attempt to do this for ethnic group differences. In reality, of course, a choice always is made, even if unknowingly, for any assessment system. Often, this choice will be disguised by an appeal to historical precedent: namely that any new test should, broadly, reflect our current knowledge about matters such as group differences. The problem is that this 'knowledge' is essentially an historical accumulation of successive decisions of the same kind, and to a large extent therefore, reflects past cultural assumptions about

such differences. Little research has been done in this area, so that we have no detailed account of how the process operates. Thus, it would be hardly surprising if the cultural assumptions and expectations of the early test constructors influenced their choice of test item contexts and hence gender differences. Gould (1981) demonstrates how a similar process influenced ethnic group differences on IQ tests.

Once such assumptions have become incorporated into existing tests it is not difficult to see how an historical determinism can be perpetuated. Unfortunately, the technical edifice which now supports the process of test construction tends to ignore such difficult issues, preferring to define the problems as technical rather than ideological or philosophical.

Essentially the same issue has been debated recently in the United States in the so called 'Golden Rule' case (Rooney, 1987). An out of court settlement in 1984 between the Golden Rule Insurance Company of Illinois, and Educational Testing Service (ETS) established the principle that, after following normal procedures for item selection in test construction, those remaining items which showed the smallest difference between blacks and whites were to be preferred.

In practice this procedure has led to difficulties in its implementation (Anrig, 1987). Nevertheless, one result has been that ETS have provided greater public access to their test construction materials and procedures. It is also interesting that the response of many psychometricians in the debate has been the standard one. Namely, to propose more refined statistical procedures for detecting statistical bias, to avoid, it seems, addressing the substantive issue which is largely political (see, for example, Anrig, 1988; Linn and Drasgow, 1987). A detailed discussion is given by Goldstein (1989).

Where the fate of individuals and institutions may depend on the results of such assessments, then equity suggests that the process of assessment design should be open to public scrutiny and debate, and it should, of course, be a debate principally about values, aims and consequences.

Conclusions

This brief review inevitably is critical. None of the documents issued by the Schools Examination and Assessment Council, nor the TGAT report itself contain serious discussion of basic assumptions and major proposals are presented with an enthusiasm and conviction quite lacking in critical awareness.

While the idea that 'standards' will be raised appears throughout official pronouncements, there is little evidence for such a claim, nor is there much concern to define what is meant by 'standards'. Yet in some of the public debate there is a realization that there are different kinds of standards and different ways of changing them, most notably by providing extra resources.

Above all, as more of the 1988 Educational Reform Act comes into operation, so it becomes clear that the extreme haste in which many of its proposals were conceived, has produced undesirable consequences. Nowhere is this more apparent than in the area of assessment where the attempt to impose an elaborate structure upon an inadequately formulated theoretical base may eventually destroy the whole edifice.

Notes

1 The grading system for GCE and CSE is on a seven-point scale for each subject. The highest grade, 'A', receives a point score of 7, 'B' a score of 6 etc. Thus, for example, a difference of twenty-one points on the scale is equivalent to three extra 'A' grades. For each child, the total score is simply the sum of the separate subject scores.

References

AITKEN, M. and LONGFORD, N., 1986, 'Statistical modelling issues in school effectiveness studies', *Journal of the Royal Statistical Society*, A, **149**, pp. 1–43.

ANRIG, G.R., 1987, 'ETS on "Golden Rule"', *Educational Measurement, Issues and Practice*, **6**, pp. 24–7.

ANRIG, G.R., 1988, 'ETS replies to Golden Rule on "Golden Rule"', *Educational Measurement, Issues and Practice*, **7**, pp. 20–21.

Assessment of Performance Unit, 1986, *A Review of Monitoring in Mathematics, 1978–1982*, London: DES.

CANNELL, J.J., 1988, 'Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average', *Educational Measurement; Issues and Practice*, **7**, pp. 5–9.

DES, 1988a, *National Curriculum: Task Group on Assessment and Testing, a report*, London: DES.

DES, 1988b, *National Curriculum: Task Group on Assessment and Testing, Three Supplementary Reports*, London: DES.

DES, 1988c, 'Kenneth Baker sets out principles for assessment and testing in schools', Press Notice, 175/88, June 7th 1988.

FITZGIBBON, C., 1990, *Performance Indicators — A dialogue*, BERA.

GOLDSTEIN, H., 1983, 'Measuring changes in educational attainment over time: Problems and possibilities', *Journal of Educational Measures*, **20**, pp. 369–77.

GOLDSTEIN, H., 1989, 'Equity in testing after Golden Rule', Paper read to American Educational Research Association meeting, San Francisco, March 1989.

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

GRAY, J. and JESSON, D., 1987, 'Exam Results and Local Authority League Tables', *Education and Training*, UK, pp. 33–41.

LINN, R.L. and DRASGOW, F., 1987, 'Implications of the Golden Rule Settlement for Test Construction', *Educational Measurement, Issues and Practice*, **6**, pp. 13–17.

NCC, 1988, *Consultation Report: Mathematics*, York: NCC.

NUTTALL, D.L., GOLDSTEIN, H., PROSSER, R. and RASBASH, J., 1989, 'Differential School Effectiveness', *International Journal of Educational Research*, **13**, pp. 769–76.

POPHAM, W.J. and HUSEK, T.R., 1969, 'Implications of criterion Referenced Measurement', *Journal of Educational Measurement*, **6**, pp. 1–9.

ROONEY, J.P., 1987, 'Golden Rule on "Golden Rule"', *Educational Measurement, Issues and Practice*, **6**, pp. 9–12.

THORNTON, G., 1986, *APU Language testing 1979–83: an independent appraisal of the findings*, London: DES.

WALKERDINE, V., 1984, 'Developmental psychology and the child-centred pedagogy: The insertion of Piaget into early education' in HENRIQUES, J. *et al.*, *Changing the Subject*, London: Methuen.

WOLF, A., 1987, 'How Generalizable are General Skills? issues is the assessment of competencies', Paper read to American Educational Research Association annual meeting, Washington, April 1987.

WOODHOUSE, G. and GOLDSTEIN, H., 1988, 'Educational performance indicators and LEA league tables', *Oxford Review of Education*, **14**, pp. 301–20.

4 Hierarchies in Mathematics: A Critique of the CSMS Study

Declan O'Reilly

Introduction

The view that mathematics is hierarchical by nature is both widespread and deep-rooted. The implications of this view for the learning of mathematics are profound. The Government through the National Curriculum has prescribed a single hierarchy for all children in State schools in England and Wales to follow. But where lies the legitimacy for this particular pathway through mathematics?

The major research effort on hierarchies in mathematics in this country has been carried out by the the Concepts in Secondary Mathematics and Science research project (CSMS). They too established a particular pathway through mathematics, one furthermore which claimed to be based on large-scale testing and rigorous research methods. Whilst the hierarchies posited by the National Curriculum working party and the CSMS differ in the contents of their levels and stages, both bodies appear to share the view of mathematics as a hierarchical subject, and there seems little doubt that the National Curriculum attainment targets derived considerable legitimacy from the work carried out by the CSMS team.

This paper, by focusing on some of the assumptions underlying the CSMS research, and by examining the methodology within it seeks to challenge hierarchical orthodoxies in the teaching of mathematics generally. It is argued that, whilst the CSMS study contains valuable information concerning the errors and strategies which children make and adopt in learning mathematics, its 'hierarchies of understanding', rather than being universal in application are *at best*