# Statistics and Educational Testing

*by Harvey Goldstein*

## 1. INTRODUCTION

The 1988 act and regulations require schools to publish their national assessment results, together with a statement from the LEA. This has been generally interpreted as following the recommendations of the TGAT report. That is, average results* for each school will be published at each age, with the possible exception of age 7, together with an explanation by the LEA of the school context. The latter might be a description of the social background of the school's pupils, although the TGAT report itself havers over this.

Little official comment has so far been available, either from DES or SEAC as to how precisely this scheme is to operate. Recently, however, both SEAC and HMI have shown concern about the feasibility of the TGAT proposals, especially in the light of recent research which throws doubt on their viability.

## 2. REPORTING SCHOOL AVERAGES

The most relevant research bearing on this is that carried out by Woodhouse and Goldstein at the Institute of Education (Oxford Review of Education, 1988, 14, 301-320). They reanalysed DES data on 16+ exam results using average exam scores for each LEA in England and Wales. Two substantial conclusions emerge from this.

1. Ranking LEA's on the basis of their overall average exam scores largely reflects the social and demographic composition of the Authority. LEA's such as Harrow and Barnet coming near the top and

*The TGAT report talks about presenting the distribution of scores or grades for each school, but exactly the same problems attend such reports. For simplicity we shall consider simple school score or grade averages.

those such as Haringey near the bottom. It is crucial, therefore, if anything is to be said about the *quality* of education that LEA's be compared fairly by taking such unequal 'starting factors' into account.

This was recognised by the DES in their initial (1983/84) analyses where they attempted to make adjustments for such factors as the proportion of one parent families and the social composition of the LEA. This, however, leads to the second issue which Woodhouse and Goldstein were principally concerned with.

2. Attempts to make adjustments between LEA's on the basis of average exam results, in order to try to make fair comparisons, is inherently unreliable. Very minor changes to the statistical 'model' produce very different final rankings. This is an inevitable consequence of using average LEA results rather than analyzing individual student data. Thus, for example, the analysis can be made to place Barnet second from the top or 35th from the top, or to place ILEA 56th from the top or 11th from the top.

   This inherent instability rules out attempts to rank LEA's, or in National Assessment, schools, on the basis of average results alone, whatever other information may be available for use in making allowance for social etc. factors. Any attempt to do so, whether by a formal statistical analysis or an informal procedure such as suggested by TGAT, is likely to result in lengthy, and ultimately inconclusive, debate between alternative and equally arbitrary analyses. Such a debate will throw little light on genuine school differences and the reasons for them.

   In particular, it will often obscure really important differences. It may not detect genuine failures in some schools and it may ignore excellence elsewhere.

## 3. MULTILEVEL MODELS

Most researchers in this area are now agreed that school (or LEA) comparisons should be based upon the analysis of individual student data concerning test or exam results or any other outcome of schooling of interest. Relevant recent studies which recognise this and use the appropriate statistical procedure known as 'multilevel modelling' are the ILEA Junior School Project (School Matters by Mortimore et al, 1988) and



the analysis of exam results in ILEA secondary schools by Nuttall and Colleagues (Differential School Effectiveness, by Nuttall et al, 1990, International J. of Educational Research, 13, 769-776). This research points to two conclusions.

1. That analyses which use individual pupil data together with information about the 'context' of learning such as school organisation, peer achievement, etc. can be very informative about the factors which influence achievement. They also point to the multidimensional nature of the effects of schools. Thus, in the ILEA exams analysis it was found that schools differed in the way the achievements of different ethnic groups were affected. They also differed in terms of how students with different initial achievements at the time of entry to school fared, and in the extent to which the achievements of boys and girls differed from school to school.
2. It is crucial that the achievements of students on entry to a school are taken account of. In order to measure the effects of the school per se, given the inevitable unequal and sometimes highly selective intakes, these achievements must be incorporated in any analysis which claims to make 'fair' comparisons. Often referred to as 'value added analysis' this is now generally accepted by researchers in this field.

While the basic requirements for a sound analysis are now recognised, we are still a long way from being able to 'prescribe' a standard analysis which can be carried out routinely in order to make definitive school comparisons. Rather, such analysis procedures that we do have available are best viewed as research tools designed to provide a better insight into factors which promote or inhibit learning and achievement. Statements about individual schools should be regarded as tentative, and the procedures regarded as screening devices which might indicate institutions which require more detailed follow-up, for example, via a detailed inspection.

## 4. WHITHER NATIONAL ASSESSMENT?

If National Assessment is not to be used for routine reporting of school differences, what functions remain? There is the diagnostic function, but this has been played down so far and there is little evidence that SATS will be very useful as diagnostic instruments. They could be modified with this

in mind, and perhaps some thought should be given to that. If that were to happen the whole exercise would be much smaller, and ultimately much more useful.

More fundamentally, however, is the underlying issue of how the national curriculum is to be monitored. It always seemed strange that the implementation of the curriculum should be monitored so indirectly by testing the students. As explained above, test scores are only partly the result of curriculum and schooling. If the implementation is to be evaluated then it is best done directly by observing its actual implementation in schools. There are a number of ways of doing this, from direct inspection to self evaluation by teachers and schools. Such an orientation has had little public discussion and could be taken as a starting point for a rationally argued alternative to the present system.

*Harvey Goldstein March 22, 1990*



# Discussion of Harvey Goldstein's "Statistics and educational testing"

Discussion focussed on critiques of, and alternatives to, large-scale "summative" testing of schools as recommended in the National Curriculum and T-GAT. Harvey recommended the critique of the Education Reform Bill by the London Diocesan Board for Schools (Sept. 1987), as more useful than that of the Labour Party.

The question was asked: why haven't U.S. researchers attacked the "league tables" in use there? Partly because testing by agencies such as the Educational Testing Service (ETS) is so central to the educational system. And partly because radical educational opposition is marginal in the U.S.

The question was asked: why do teachers allow themselves to be harassed by the government on issues like the national curriculum and national testing? A relevant question for those of us who work in any branch of public service—sorry, I mean, of marketing suitable service-type commodities in the market.

1992 will bring pressure to develop educational indicators—not only test scores, but also indicators about catchment areas, e.g. language difficulties, special needs. There is already a great deal of interest in performance indicators in the U.K., with the recent discussions about "Data Envelopment Analysis" (see the appendix to Woodhouse and Goldstein, 1988, for a critique); DEA seeks to measure the efficiency of LEAs in producing a range of educational outcomes against others with similar characteristics. And the DES is currently seeking advice in Polytechnics about suitable choices for performance indicators.

*Jeff Evans*

# Harvey adds a later statement about performance indicators

There is a great deal of interest internationally in so-called 'performance indicators'. These are essentially aggregate level descriptors or indicators of institutional characteristics. Inevitably, most interest tends to centre on 'output' measures such as exam or test scores. Analysis, presentation and interpretation of such measures is explicitly in terms of such aggregates and this results in serious problems. Woodhouse and Goldstein (1988) pointed out and demonstrated with DES exam data the inconsistency of modelling at school or LEA aggregate level. They pointed out that 'explanatory' models were self contradictory since they attempted to rank institutions on the basis of model residuals while simultaneously trying to produce models in which the statistical assumption of random residual variation was approximated! As would be expected, such procedures in fact produce highly unstable results against minor model perturbations. They also pointed out that ALL techniques based upon such aggregate level analyses, such as data envelopment analysis suffered from the same drawbacks. One conclusion is that while performance indicators might play a useful role in describing the state of institutions they are not appropriate for exploring relationships between factors with a view to making inferential statements about influential factors. Thus, for example, reporting the level of resources in a school may be useful in the context of a policy of equitable resource allocation, but relating schools' resources to average examination results tells us nothing of interest about the effect of one upon the other. To do that we would require data on exam results for each student and the way in which each student had access to resources.

*Harvey Goldstein April 26, 1990*