*Cambridge Journal of Education*
Vol. 38, No. 3, September 2008, 393–400



# Evidence and education policy – some reflections and allegations[1]

Harvey Goldstein*

*Graduate School of Education, University of Bristol, Bristol, UK*

*(Received 13 September 2007; final version received 16 January 2008)*

The paper reflects on the use by the UK central government of statistical evidence in educational policy matters. Particular attention is given to school league tables. The paper is generally critical of government attitudes, but suggests that progress towards rational decision-making does occur.

**Keywords:** evidence-based policy; league tables; education policy



## Attitudes to evidence

Within the UK the New Labour administrations of Tony Blair and Gordon Brown have claimed that wherever possible their policies are based upon the best research evidence. In this paper I shall attempt to explore the basis of this claim insofar as the use of statistical information is concerned, with particular reference to the history of performance indicators (PIs) in the form of school league tables. While I shall concentrate on the policies and actions of the present government and its advisors, I have few reasons to believe that they differ in kind or intention from the policies of previous (and indeed future) UK governments. Indeed, among the examples I shall discuss there are issues that resonate with other educational systems. In particular, the 'high stakes' publication of school rankings, or league tables, is a matter of considerable debate within the educational systems of countries such as Australia, the USA and Ireland, and I would expect that observers of those systems will recognise many of the issues I shall be discussing.

By the middle of the first decade of the twenty-first century a somewhat critical, even cynical, attitude towards government use of evidence could be found among researchers. Thus, Michael Rutter, a highly respected leading social researcher in the UK, remarked that: 'The Government definitely does not want evidence, although their rhetoric is entirely different … . They just care about it if it fits their plans' (quoted in Shepherd, 2006). A well-known educational researcher, Peter Tymms, claims that 'There was a coordinated attack and a rubbishing of my research [on homework] by the Prime Minister, David Blunkett [then Secretary of State for Education] and Chris Woodhead (head of the Office for Standards in Education at the time)' [OfSTED, which is the body charged by the Government to carry out inspections and publish reports on schools] (quoted in Shepherd, 2006). Interestingly, the same article in the *Times Higher Education Supplement* claimed that some 80% of academics reported that they could no longer properly convey what they felt was good research evidence to policy-makers where the results might be unwelcome.

*Email: h.goldstein@bristol.ac.uk

ISSN 0305-764X print/ISSN 1469-3577 online

DOI: 10.1080/03057640802299643
http://www.informaworld.com

The allegation in the first quotation, that the Government uses only that evidence that suits its preconceived views, is arguably one that could be applied to many governments in many places at many different times. It is, of course, possible to have some sympathy with governments here. It may well be the case that when research evidence is set alongside other information and contextualised in terms of cost, feasibility and even public acceptance it will be rational to reject its (implied) recommendations. Rutter, however, is making a stronger point, namely that the government is pretending to use research evidence in an appropriate fashion, but in fact simply ignores it whenever the research results are inconvenient, rather than honestly trying to incorporate its findings into its policies. Thus, what is different about the present New Labour government, as opposed to previous administrations, is the repeated emphasis that it has given to evidence-based policy in speeches and other pronouncements, especially in the area of education and social policy. For example, the 1999 White Paper *Modernising government* stated that:

> policy decisions should be based on sound evidence. The raw ingredient of evidence is information. Good quality policy making depends on high quality information, derived from a variety of sources – expert knowledge; existing domestic and international research; existing statistics; stakeholder consultation; evaluation of previous policies. (Cabinet Office, 1999, p. 31)

Blunkett himself argued for the need for sound evidence and that it is 'self-evident that decisions on Government policy ought to be informed by sound evidence' (Blunkett, 2000). If Rutter is right, however, this emphasis on evidence, and the associated resources given by the government to 'policy related research' can legitimately be viewed simply as another example of 'spin' intended to legitimise that policy in the name of scientific research.

The allegation in the second quotation is somewhat different, and if justified seems more sinister since government is here accused of using its power and influence to undermine the integrity of particular researchers. There are effectively two issues. The first concerns the quality of the research itself. Thus, the research referred to reported counter-intuitive results that many would have found surprising.[2] In such a situation those civil servants and others who were advising the politicians would certainly have suggested caution, and perhaps advised seeking further opinions or research. I shall return to this below. What is new, however, and quite unethical, is the intervention of politicians, with no personal competence in the area, directly in a research debate. I had a similar, but much less traumatic, experience when Kenneth Clarke, as secretary of state for education during a previous Conservative administration, poked fun at myself and a colleague in public by deliberately transposing the syllables of our surnames. We had published research which presented arguments against the publication of school league tables, which was then a relatively new feature of government policy (see Nuttall, Goldstein, Prosser, & Rasbash, 1989).

This leaves open the issue of research quality and how any given piece of research should be evaluated. I include in this so-called 'systematic reviews' of research which, like any individual research project, are contingent on those carrying out the review, their prior experiences and assumptions. This is a difficult problem and raises all the familiar epistemological issues with which researchers have to grapple. The point I want to make is that researchers need to be quite careful about who they blame for any misinterpretation. Let me illustrate this by looking more closely at the Tymms'

research on homework. This was published shortly after a previous report partly authored by Blair's then principal advisor on education, Michael Barber, which concluded that homework was associated with improved performance and lent support to the Labour Party's current policy (January 1997) in favour of mandatory periods of homework (Barber, Myers, Denning, Graham, & Johnson, 1997). In fact that research had serious flaws (see http://www.cmm.bristol.ac.uk/team/HG_Personal/home_rep.html for a critique) but did seem to resonate with received opinion. The Tymms' report was, therefore, very 'high stakes'. Unfortunately, the political debate that ensued effectively prevented a proper peer review of the report itself as well as the currently available evidence. In fact, the Tymms' report, which claimed that large amounts of homework were associated with lower performance, provides poor evidence. While the analysis did adjust for school differences and factors such as verbal reasoning scores, it failed to take account of prior achievements in the curriculum subjects under investigation. Thus, the sample it used was effectively cross-sectional, so that there was no way of telling whether previous poor performance was responsible for pupils doing more homework or whether actually doing large amounts of homework depressed performance. There is perhaps a certain irony in that had the politicians acted responsibly, they may well have been able to substantiate their policies through critical peer review of the Tymms' research, which could have provided a more secure basis for their own homework policies. There is also, therefore, an important lesson for politicians, namely that acting from the best, rather than basest, motives may actually, even in the short term, be to their advantage!

Before turning to league tables let me mention another issue that for many years has been a major concern of policy-makers and educationists, namely the effect of class size on learning. Despite the common assumption that lowering class sizes results in improved learning, research evidence, despite thousands of studies until the late 1980s, was largely lacking. A meta analysis (Goldstein, Yang, Omar, Turner, & Thompson, 2000) has shown that by the start of the twenty-first century there were just nine international studies of primary school age children that satisfied the basic criteria for providing sound evidence, principally the requirement to have longitudinal data. The analysis showed that there was indeed a modest improvement in achievement (about 0.2 standard deviations for a reduction of 10 pupils). These studies and a more recent one (Blatchford, Goldstein, Martin, & Browne, 2002) also found that the effect of class size largely operated in the first (reception) year of formal schooling and seemed to be greater for more disadvantaged children. Interestingly, while the government does now have a policy for all infant classes (one teacher should not teach classes greater than 30) the findings concerning the importance of the reception year and on differential effects have not filtered through to policy.

Some of the opponents of policies aimed at reducing class size, prior to the more recent evidence, did use the lack of association to argue against class size reductions. Nevertheless, it can be argued that because of the prima facie plausibility of an association between class size and performance it was entirely justified for researchers and funders of research to continue to study the issue, using improved techniques. This policy was vindicated when the STAR study (Finn & Achilles, 1999), a large randomised trial, found an association in the early years and, following that breakthrough, further studies have confirmed the finding. In other words, all research evidence is provisional, especially in the social sciences, and

forming judgements is often rather more complicated than simply appealing to 'evidence'.

An informative example of the tentative nature of research evidence when related to policy is the case of the International Adult Literacy Survey (IALS). In the early 1990s France pulled out when it emerged that the results for literacy levels in France were unexpectedly low. Here, although the French unwillingness to confront unwelcome evidence may have been more influenced initially by politics than science, a perfectly reasonable scientific case for scepticism could have been made, and indeed was made in order to undertake a re-analysis. When research evidence seems to be counter-intuitive or to contradict other evidence it is perfectly reasonable to seek further evidence or debate in order to provide an explanation. In fact, the subsequent re-analysis exposed a variety of flaws in the design and execution of the study that not only supported the French stance but led to useful insights into the nature of such international comparative research (Blum, Goldstein, & Guerin-Pace, 2001).



**League tables – a short history of nearly everything a politician shouldn't do**

It was in 1986 that the administration of prime minister Margaret Thatcher, building upon work carried out by the Inner London Education Authority,[3] first tentatively decided to publish secondary school average examination results and thus provided the means for ranking schools on what were claimed to be measures of school 'quality'. This policy was strengthened over the next few years. During this time the UK government introduced 'Key Stage tests' at the ages of 7, 11 and 14, and by the time of the New Labour government in 1997 the 11-year-old (Key Stage 2) test results were also being published. Parents were encouraged to use the rankings in their choice of schools. The Royal Statistical Society report on performance indicators (http://www.rss.org.uk) provides a broad review of the issues surrounding the use of league tables in a number of areas, including education, health and crime. A technical discussion of the statistical issues has been given by Goldstein and Spiegelhalter (1996). Briefly, the main issues are as follows.

The first rankings to be published were simply percentages of pupils in each school achieving the highest grades in the GCSE and A-level examinations,[4] these being the certification examinations generally taken at ages 16 and 18, respectively. A scoring system based upon all the examination grades was also used with the rankings, based upon the average score for each school. From the outset many in education had pointed out the shortcomings of the league table policy, citing research findings that demonstrated the need to adjust results for school intake characteristics (the value added model) and also the need to provide uncertainty (confidence) intervals[5] for the mean scores based on relatively small sample sizes. Nuttall et al. (1989) provided an early critique that demonstrates the inadequacy of these rankings by using research data where intake measures were available. They showed that after adjustment the rankings of many schools were changed and that when confidence intervals were placed around the school estimates most schools could not be statistically distinguished from the average. This was later reinforced by Fitz-Gibbon (1997) in an officially commissioned report.

In response the government, quite properly, was able to point out in the early days that the data were not available to carry out such adjustments for the whole

population, and indeed these data have only really become available in the last five years or so. Nevertheless, in 1995 the then Conservative government first officially committed itself to value added 'performance tables' and by 2007, thanks partly to the existence of the National Pupil Database (http://www.bris.ac.uk/Depts/CMPO/PLUG/) containing longitudinal pupil data, these have become regular publications. Still, there remains a considerable reluctance to embrace the notion of confidence intervals, as I shall elaborate below.

There is a whole set of key issues about the ways in which the imposition of a 'high stakes' national testing regime affects the behaviour of schools, pupils and parents. These include incentives by all players to maximise their test scores at the expense of longer term learning goals; the stress that testing imposes upon pupils; the encouragement of 'gaming' by institutions, all the way up to outright cheating. In addition, the use of test scores for accountability suffers from the major (and inevitable) drawback that the results being used apply to a cohort entering the institutions several years earlier – up to six years in the case of GCSE examination results. We know that the (adjusted) correlations across such long time intervals are only moderate, i.e. of the order of 0.5 (Gray, Goldstein, & Thomas, 2001). As a result, any utility such comparisons might otherwise have is substantially reduced. While all these issues are important I shall not dwell further on them – rather I shall concentrate on the statistical issues of adjustment and uncertainty.

It is clear that the UK government has taken some note of research evidence and the current official 'contextual value added' (CVA) performance tables go some way to meeting the technical concerns. These were first introduced in 2002 to adjust the GCSE examination results for prior attainment at the ages of 14 and then at 11 years (see http://www.standards.dfes.gov.uk/performance/1316367/CVAinPAT2005/). They also take account of other background factors, such as ethnicity and free school meals eligibility, and include, for each institution, a 95% estimated confidence interval. Nevertheless, there remain considerable problems.

First, and perhaps most importantly, the media have not taken up the message about confidence intervals. When secondary school tables were published on 11 January 2007 none of the four major UK 'broadsheets' (*The Guardian*, *The Independent*, *The Times* and *The Daily Telegraph*) gave any indication that these intervals exist or are important. Indeed, *The Daily Telegraph* went out of its way to quote the chair of the 'Specialist Schools and Academies Trust' – clearly someone with a vested interest – who attempted to rubbish the CVA tables with a reference to (adjusted) results being 'manipulated'. An honourable exception was the BBC web site (http://news.bbc.co.uk/1/hi/education/6251587.stm), which had a balanced account of the relevant issues.

Secondly, the government continues to publish the unadjusted ('raw') performance tables – although this applies strictly only to England, since the other UK education systems have for some time ceased to publish league tables. Moreover, the government has also consistently failed to recognise the need for confidence intervals for these, despite providing them for the CVA tables! It is not entirely clear how they justify such inconsistency, but they do claim that to provide intervals for raw school results would be 'potentially confusing for schools' (letter from the chair of the Statistics Commission to Harvey Goldstein, 23 May 2005). It would seem that they also claim to believe that since the school averages are based upon the whole 'population' of pupils statistical uncertainty estimates are unnecessary – a view that

seems to be derived at least in part from a serious ignorance of statistical inference as applied to social processes, a view propagated even in certain academic quarters (for a discussion see Goldstein & Noden, 2004).

Thirdly, if the government insists that, in the words of the then Secretary of State Alan Johnson, the league tables are 'non-negotiable' (quoted by James Meikle, education correspondent, *The Guardian*, 9 January 2007), then it is passing up an excellent opportunity to explain to the public some important aspects of statistical reasoning concerned with adjustment factors, sampling and uncertainty estimates. Instead, its current poorly informed policies seem more likely to enhance the public's suspicions about government statistics than to offer reassurance.

Finally, to be fair, the government's DCFS website does have a reference to a departmental paper that does a reasonably competent job in explaining the technicalities (http://www.standards.dcsf.gov.uk/performance/1316367/CVAinPAT2005/?version=1). In fact, an earlier version of the website cited three papers purporting to explain the CVA methodology, but that section of the website was removed with the eventual intention of presenting a revised version. It seems that this was in response to the presentation of an earlier version of the present paper at the Royal Statistical Society conference in York in July 2007, and is a welcome sign that the government can be responsive to critical comment. Of the three papers which are still available one is an unpublished conference presentation (http://www.leeds.ac.uk/educol/documents/143649.htm) that in fact makes no reference to the extensive literature on the topic but just happens to argue against any kind of value added measure. The second paper (http://www.lums.lancs.ac.uk/news/7002/) echoes the argument against raw league tables, but clearly fails to understand what CVA really is. The third paper actually refers to an NFER review which contains little discussion of the issues.

The league table culture is symptomatic of a deeper problem with public debate that should concern citizens. Namely, a surface precision associated with numerical data is used, sometimes unscrupulously, sometimes in ignorance, as a substitute for serious and well-informed debate. The promotion of school league tables as if they convey uncontestable information about schools is just one example and it is rather worrying that such tables are now being introduced into higher education, for example in the new student ratings of teaching on undergraduate courses (http://www.cmm.bris.ac.uk/team/HG_Personal/hefce-student-pilot-commentary.htm). Here, however, under the auspices of the Higher Education Funding Council (HEFCE), there is a more enlightened view about the need to provide confidence intervals and, in particular, there is a concern about optimum ways of presenting such intervals so that their properties are well understood by the general public.

## Some conclusions

It should be fairly clear that my view about the use of evidence in education policy-making is fairly pessimistic. It does seem to be the case that when sufficient people with obvious expertise take up an issue over a period of time then government does listen. Whether it does so because it senses that it might otherwise lose important electoral support or whether it has a genuine interest in promoting rationality and public understanding is an interesting question, but beyond the scope of the present paper. However, it is certainly the case that in other areas of public policy, notably most recently over the decision to invade Iraq, the present government is clearly

prepared to act in defiance of the evidence, and even in the knowledge that it may well lose public support as a result.

It is also important to recognise the significance of the media – especially the mass circulation press. All governments are to some extent fearful of what these media will say and they may often take the view that placating the likes of the mass circulation *Sun* and *Daily Mail* newspapers is itself a good way to electoral success. As I have pointed out, however, even the so called 'quality' press leaves a great deal to be desired in terms of understanding and serious presentation.

I have also argued that researchers themselves are not beyond reproach. Not all research, wherever published, is necessarily of high quality. As in all areas of scientific activity there is no proper substitute for continuing debate and discussion among peers, and a commitment to such discussions before adopting policies would be a welcome move from government.

Nevertheless, in the present climate, on the evidence we currently have about government attitudes towards the use of research evidence, perhaps the most fruitful approach for those concerned with genuine progress towards evidence-based policy is not to concentrate on trying to persuade policy-makers. That way will often lead to frustration, or even to annihilating cynicism. Rather, it may be more rewarding to concentrate on public and professional education. The former is difficult, for all the usual reasons to do with limited access to the media, and it remains an urgent task for researchers and professional organisations to seek ways to promote public education, whether this is directed towards other education professionals such as teachers and academics, towards government statisticians or simply towards interested members of the public.



## Notes

1. A version of this paper was read to the Royal Statistical Society annual conference, York, 2007.
2. The research claimed that among 11-year-olds there was a negative association between the amount of homework done and educational attainment. (Farrow, Tymms, & Henderson, 1999).
3. This was the body that organised primary and secondary education for inner London up to its disbanding by the Government in 1990 and the allocation of education management to the component local authorities in the area.
4. The GCSE is the General Certificate of Secondary Education taken by 16-year-olds at the end of compulsory schooling in Year (grade) 11. The A-level is the advanced level General Certificate of Education examination that is taken at the end of Year 13 and principally serves as a university entrance qualifying examination.
5. A confidence interval provides a range of values that, with a given probability – typically 0.95 – is estimated to contain the true value of the school score. If such an interval includes the value zero then an equivalent statement can be made that the true value is not significantly different from zero at the 5% level.

## References

Barber, M., Myers, K., Denning, T., Graham, J., & Johnson, M. (1977). *School performance and extra-curricular provision*. London: Department of Education and Employment.

Blatchford, P., Goldstein, H., Martin, C., & Browne, W. (2002). A study of class size effects in English school reception year classes. *British Educational Research Journal*, *28*, 169–185.

Blum, A., Goldstein, H., & Guerin-Pace, F. (2001). International Adult Literacy Survey (IALS): An analysis of international comparisons of adult literacy. *Assessment in Education*, *8*, 225–246.

Blunkett, D. (2000). *Influence or irrelevance: Can social science improve government?* Swindon: ESRC.

Cabinet Office. (1999). *Modernising government*, White Paper CM4310. London: The Stationary Office.

Farrow, S., Tymms, P., & Henderson, B. (1999). Homework and attainment in primary schools. *British Educational Research Journal*, *25*, 323–342.

Finn, J.D., & Achilles, C.M. (1999). Tennessee's class size study: Findings, implications, misconceptions. *Educational Evaluation and Policy Analysis*, *21*, 97–109.

Fitz-Gibbon, C. (1997). *The Value Added National Project final report – feasibility studies for a national system of value-added indicators*. London: School Curriculum and Assessment Authority.

Goldstein, H., & Noden, P. (2004). A response to Gorard on social segregation. *Oxford Review of Education*, *30*, 441–442.

Goldstein, H., & Spiegelhalter, D.J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society*, *159A*, 385–443.

Goldstein, H., Yang, M., Omar, R., Turner, R., & Thompson, S.G. (2000). Meta analysis using multilevel models with an application to the study of class size effects. *Journal of the Royal Statistical Society*, *49C*, 399–412.

Gray, J., Goldstein, H., & Thomas, S. (2001). Predicting the future: The role of past performance in determining trends in institutional effectiveness at a level. *British Educational Research Journal*, *27*, 391–406.

Nuttall, D.L., Goldstein, H., Prosser, R., & Rasbash, J. (1989). Differential school effectiveness. *International Journal of Educational Research*, *13*, 769–776.

Shepherd, J. (2006, December 1). 'I felt very isolated,' says stunned critic. *Times Higher Education Supplement*.