'This is an electronic post-print version of an article published in *Educational Review* Vol. 60, No. 2 (May 2008): 179-185. *Journal of Education Policy* is available online at: http://www.tandf.co.uk/journals/titles/0013-1911.asp.

URL to published version: http://dx.doi.org/10.1080/00131910801934185.

The value-added of primary schools: what is it really measuring?

Stephen Gorard
School of Education
University of Birmingham
B15 2TT
s.gorard@bham.ac.uk

Abstract

This paper compares the official value-added scores in 2005 for all primary schools in three adjacent LEAs in England with the raw-score Key Stage 2 results for the same schools. The correlation coefficient for the raw- and value-added scores of these 457 schools is around +0.75. Scatterplots show that there are no low attaining schools with average or higher value-added, and no high attaining schools with below average value-added. At least some of the remaining scatter is explained by the small size of some schools. Although some relationship between these measures is to be expected – so that schools adding considerable value would tend to have high examination outcome scores – the relationship shown is too strong for this explanation to be considered sufficient. Value-added analysis is intended to remove the link between a schools' intake scores and their raw-score outcomes at KS2. It should lead to an estimate of the differential progress made by pupils, assessed *between* schools. In fact, however, the relationship between value-added and raw scores is of the same size as the original relationship between intake scores and raw-scores that the value-added is intended to overcome. Therefore, however appealing the calculation of value-added figures is, their development is still at the stage where they are not ready to move from being a research tool to an instrument of judgement on schools. Such figures may mislead parents, governors and teachers and, even more importantly, they are being used in England by OFSTED to pre-determine the results of school inspections.

Introduction

Much has been written about the problems involved in making comparative claims about the relative effectiveness of schools with equivalent pupils (Gorard 2000, 2001, 2005). There are difficulties in assuming that the indicators of school outcomes are comparable across time, place and curriculum area. There are also difficulties in equating outcomes scores for pupils at one age with scores at a later age. These

difficulties are exacerbated by the limitations of the methods used to address comparability. In general, education analysts do not conduct 'active' studies that involve allocating pupils to schools, teachers, or examinations for research purposes. For a variety of practical and ethical reasons, analysts find themselves faced with the rather more 'passive' analysis of datasets, over which they have no control. The problem with this 'post hoc dredging of sullen datasets' (Gorard 2006a) is that the statistical methods usually involved were designed for use only in active research (Lunt 2004).

The design of experimental approaches to research allows us to make observations of difference or pattern in practice that can be directly related to a prior theory or hypothesis (Gorard 2002). A problem arises, however, when this logic of experimentation is extended to other approaches, such as the regression analyses used to create value-added measures (Gorard 2006b). Without a controlled trial, the direct link between a hypothesis and its testing in practice disappears, and is replaced by a much a weaker form of 'test', such as those based on probability and significance. The results of these can be very misleading (Lunt 2004). For, in most research situations, it is not sampling variation that is the key to understanding and unlocking the process (Ziliak and McCloskey 2004). However, sampling variation is all that traditional statistical analysis addresses, and often not very well at that (Gigerenzer 2004). Researchers should be more concerned with developing and using indicators of the scientific importance of their results, than with how well the results fit to a rather arbitrary statistical model. For example, they could ask whether what they have found fits observations elsewhere, can be uncovered using a variety of different methods, whether it looks right in practice, or what the dangers might be in assuming that it is true.

This paper illustrates these points – especially the need to be sceptical about results that depend on only one method – with an important topical example. The 'raw' examination scores produced in different schools are not so much a measure of the impact of the schools as of the ability and outcome scores of their allotted pupils. In order to decide which schools are making differential progress with their pupils, the DfES in England is now producing value-added scores for each school. These value-added scores attempt to measure the differential progress made by strictly *equivalent* pupils in different schools.

Methods

In this 'value-added' analysis, the prior attainment of each pupil is taken into account, such that the published official figures reflect not the intake to the school but the average progress made by pupils while in the school. The DfES value-added scores for the average pupil progress from Key Stage 1 (KS1, the prior attainment of the pupil aged 7 at primary school) to Key Stage 2 (attainment at age 11) in each secondary school are calculated as follows (fuller details are available at DfES 2006).

Most independent schools, infant-only schools, pupil referral units and schools with less than five pupils in the age group are excluded. Otherwise, for the 2005 figures, all pupils in an eligible school were included who were eligible for KS2, still on the school roll in May 2005, and with a matched KS1 score. Each pupil KS1 and KS2 outcome

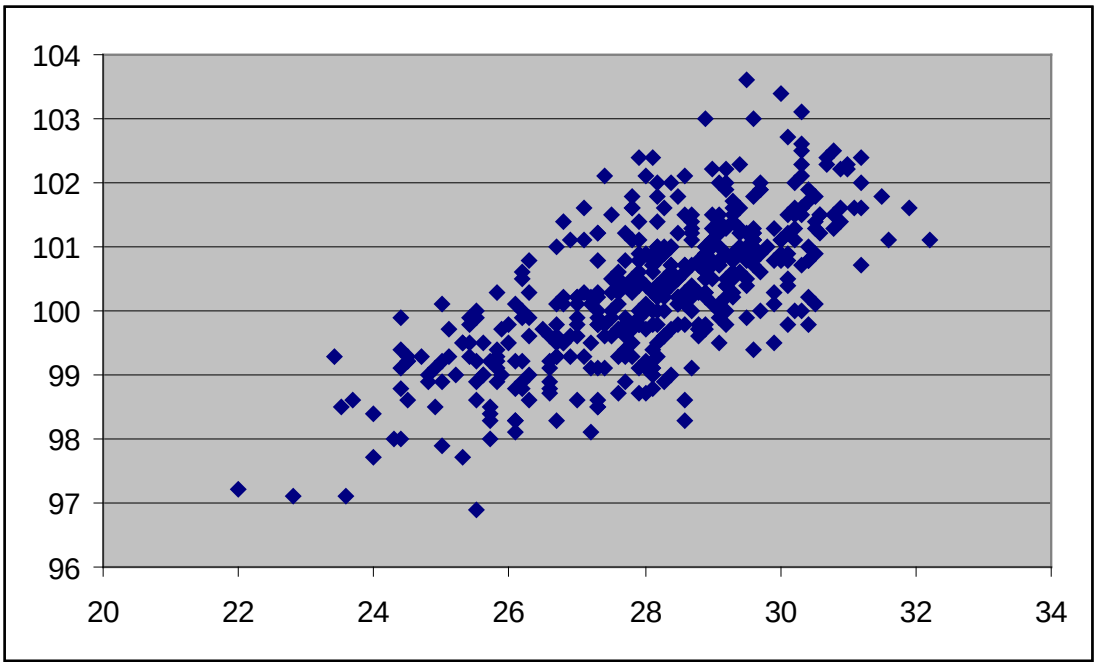
was awarded a point score (so that working towards a level 1 at KS1 is awarded 3 points, and level 4 or more converts to 27 points). The scores for overall reading, writing and mathematics at KS1, and the scores for English, mathematics and science at KS2, were then averaged for each pupil. A pupil's value-added score is calculated by comparing their KS2 average with the median KS2 score for all pupils with the same KS1 score. Thus, in mainstream schools, a pupil with an average of level 1 at KS1 (9 points) might be expected to attain an average of level 3 at KS2 (21 points), for example. The value-added score for each school is the average of the value-added scores for all pupils meeting the definition above (with 100 added to this average to eliminate negative values). 100 is near par, and a value-added score between 99.4 and 101.2 is described as 'broadly average'.

This paper uses the Key Stage 2 (KS2) results for mainstream primary schools in England in 2005, and their published DfES value-added scores. The re-analysis presented here is based on all 457 primary schools with complete information in York, Leeds, and North Yorkshire. Results are presented in scatterplot form, or as Pearson R correlation co-efficients - which can be squared to give an 'effect' size. The approach is very similar to that used in Gorard (2006c), which demonstrated that value-added scores in secondary schools in England are no more independent of raw-scores than the raw-scores are independent of the schools' intake values. Paterson (1997) found similarly high correlations between raw scores and the results of regression analyses based on pupils' prior qualification.

The same correlation appears

Figure 1 shows that the same relationship, previously noted in the DfES value-added figures for secondary schools in England and by Paterson (1997) in Scotland, also appears in the DfES value-added figures for primary schools. There is a very clear quasi-linear relationship between the KS2 raw-score for any primary school and its eventual value-added score. All of the high value-added schools (e.g. above 101) have relatively high raw scores (e.g. around 27 points or above). All of the low value-added schools (e.g. below 99) have relatively low raw scores (e.g. below 29 points).

Figure 1 – Crossplot of value-added scores against Key Stage 2 results, 457 primary schools, 2005



Note: The graph shows the DfES KS1-KS2 value-added scores, and the average points score per pupil at Key Stage 2 for all primary schools in York, Leeds, and North Yorkshire LEAs.

Both the KS1 and KS2 scores, of course, contain a considerable but unknown element of error. The Key Stage tests may have less than perfect validity in what they purportedly measure, candidates may make untypical mistakes in responding to questions, some teachers and schools may condone 'sharp practice' in administering the tests, some candidates will be missing, and some candidates will have missing scores. There may be mistakes in the marking, recording and computing of the KS2 points per school. The marking is to a threshold in which the achievement of two pupils just above and below a threshold may actually be closer than the achievement of two pupils awarded the same grade. The grades are converted to a points score, which changes the metric and may create additional distortions in the data. The value-added scores are then created from these two imperfect sets of figures, and the value-added model is only one of many possible, requiring a number of untestable analytical assumptions based on subjective judgements. This level of uncertainty in the result could be sufficient to explain the apparent differences between the value-added scores of schools with similar raw-scores in Figure 1.

The correlation between the primary schools' value-added score and their KS2 results is +0.74 (Pearson's R). One of the major reasons why this correlation is lower than that previously published for secondary schools (Gorard 2006c) is that primary schools are generally much smaller, with fewer pupils in each cohort. Therefore, there is more volatility in the figures (or put another way, the measurement problems outlined above are more apparent – see also Tymms and Dean 2004). One way of assessing whether this is the correct interpretation is to examine the correlation for large and small primary schools separately. If the correlation is lower for small schools but larger for large schools then this is an indication that the volatility of small schools helps makes the correlation 'appear' smaller than it is at the secondary level.

This *is* what happens. The correlation between primary schools' value-added score and their KS2 results drops to +0.69 for the 353 schools with less than 50 pupils in the cohort, and to +0.67 for the 255 schools with less than 32 pupils, for example. The correlation rises to +0.76 for the 354 schools with more than 18 pupils, and to +0.78 for the 258 schools with more than 30 pupils, for example. All of the schools with 50 or more pupils in the cohort had value-added scores in the narrow range of 98 to 102 and, in general, the schools with the most extreme value-added scores had very few pupils. All of this suggests that the school-level value-added scores can be explained to a large extent by the actual level of attainment of pupils at KS2 (i.e. the raw scores), and the apparent differences (the width of the scatter in Figure 1) can be explained by measurement error and the volatility of small numbers.[1]

Some commentators might suggest that Figure 1 actually shows two different kinds of regression. In addition to the bottom-left to top-right pattern discussed so far, there is also a sequence of top-left to bottom-right 'lines'. But there is no way of distinguishing such a conceptual sequence from the scatter and volatility described above. The appearance of the graph itself is affected by the scale chosen, and a visual comparison of the two kinds of slopes is, therefore, not a reliable guide to the overall pattern. If the pattern in Figure 1 had been close to a perfect diamond shape with corners at (27, 103), (27, 97), (23, 100), and (31, 100) then the correlation would be zero, or very close to zero. If, on the other hand, there had been an appreciable negative slope then the correlation would have been negative overall. But +0.74 is a very high correlation – considerably higher than standard in the educational literature – representing an 'effect' size of 55%. It is also a positive correlation, representing the positive left-right slope while ignoring the negative one.

Discussion

If accepted, then the re-analysis above, coupled with the similar analysis of the results for all secondary schools in England (Gorard 2006c), suggests two important kinds of conclusion. The first kind of conclusion that can be drawn is methodological. Many analysts agree that value-added comparisons of the kind conducted so far by DfES are problematic (e.g. Tymms and Dean 2004, Schagen 2006). Their usual response is to try and make this complex analysis even more complex. But without confirmatory evidence of a different nature, and no sceptical consideration of the meaning of the measures involved, there is a danger that large-scale complex analyses such as those considered here are rhetorically misleading. In fact, the changes and differences identified as school effects may be largely chance processes, with a greater random element than traditional analysts allow (Pugh and Mangan 2003). How do we know that the variation in value-added scores for different schools *means* anything at all? There is no external standard or arbiter to which we can refer. The calculations look plausible enough, but no one had predicted the level of correlation found between raw- and value-added scores. In fact, on hearing of it, commentators at the DfES first denied the correlation, and then attributed it to some peculiarity of schools in Yorkshire, briefing the education minister in the House of Lords to state this in

[1] In fact, since a value-added score is, in essence, the difference between prior and subsequent attainment figures, one would expect around half of the variance in VA figures to be explained by either of these raw-scores, leading to two correlations of around 0.7 each.

response to a query by another member (in Hansard – see http://www.publications.parliament.uk/pa/ld200506/ldhansrd/vo050620/text/50620w02.htm). On being shown that the same relationship held for all secondary schools in England, the reply by Lord Adonis was not re-addressed, and the findings were simply ignored.

Making the official value-added analysis more complex, via the addition of contextual information about the pupils, may disguise but will not solve the problem highlighted in this paper. The additional complexity will reduce further the number of potential critics able to understand the methods. The use of additional information about the social background of pupils is likely to decrease the scatter shown in Figure 1, making the relationship between school intakes and school outcomes stronger. The inclusion of social background information in school performance figures will also have the unintended consequence that we will no longer be able to consider the extent to which schools do, or do not, compensate for differences in those backgrounds. At present, if VA worked, we could see whether schools were equally effective for rich and poor pupils, by disaggregating the VA by eligibility for free school meals (FSM), for example. Using contextualised VA with FSM factored into the calculation, it does not make sense any longer to disaggregate by FSM. And the same point can be made about ethnicity, language, and special need. Plans to make value-added analysis more complex, through the use of advanced regression techniques will also be counterproductive. They will also reduce further the number of potential critics able to understand the methods, but can not overcome the problem of correlation noted here.

The percentage variation at school level, usually termed the 'school effect', is small and suggests incorrectly that schools are making little difference to their pupils. There is also the confusing situation that the same school may appear to be effective on one measure (such as attainment) but not another (such as dropout), or effective for one age group and not another. Therefore policies based on VA results and designed to improve test performance for one age group can hurt performance in other areas (Rumberger and Palardy 2005). The solution to all of these issues is not a more complex value-added analysis. The solution lies in re-thinking what it is that we want value-added analysis to achieve. There *are* simpler and more scientific alternatives to measuring the impact of schools. One alternative suggested recently requires no consideration of prior attainment or contextual variables, relying instead on the discontinuity of school years or grades to estimate the absolute effect of going to a school in comparison to not going to school at all (Luyten 2006).

The second kind of conclusion from this paper is more practical. Until concerns about value-added analyses have been resolved, it is not reasonable to use them for practical purposes.[2] Parents cannot rely on them when choosing schools. School leaders cannot rely on them to judge the effectiveness of teachers or departments, and officials cannot rely on them to make decisions about the quality of education delivered in schools. Rather, what this re-analysis shows is that schools with a low-attaining pupil

[2] Clearly, there will never be an ideal measure able perfectly to summarise the performance of a school. That is not the point. If accepted, what this paper shows is the DfES approach is nothing like a solution to the problem of measuring pupil progress independently of their raw-score attainment. It is neither good enough, nor even the best approach currently available.

intake have, *ceteris paribus*, low raw-scores at KS2, and that the 'value-added' scores do almost nothing to overcome this clear pattern. Therefore, these value-added scores are not, as the DfES has claimed, independent of actual levels of raw-score attainment.

An example of why this matters comes from the revised OFSTED light-touch school inspections in England. Inspectors from OFSTED are spending less time in schools, and making fewer lesson observations, on each inspection. The reduced reliance on primary observation has, in the reports of some school leaders, led to an increased reliance on prior value-added analyses of the schools (Bald 2006). This has led to clear anomalies and 'bizarre judgements' such as a school being judged largely 'good' or 'outstanding' on observation, but being reported as merely 'satisfactory' because the best outcome allowable by OFSTED was constrained by a relatively low prior value-added score (Mansell 2006a, Slater 2006). The increased reliance on contextualised value-added (at time of writing), which is very sensitive to exclusions for example, leads to some schools getting lower than expected inspection results (Mansell 2006b). As Paterson pointed out as early as 1997, pupil-level regression of the kind now in use by the DfES is a fascinating and productive research tool, which can be used to inform professional debate. But it should not yet be used directly as a tool for pupil, teacher or school assessment.

References

Bald, J. (2006) Inspection is now just a numbers game, *Times Educational Supplement*, 26/5/06, p.21

DfES (2006) *Value-added technical information*, http://www.dfes.gov.uk/performancetables/primary_03/p5.shtml, accessed 15th March 2006

Gigerenzer, G. (2004) Mindless statistics, *American Economic Review*, 33, 5, 587-606

Gorard, S. (2000) 'Underachievement' is still an ugly word: reconsidering the relative effectiveness of schools in England and Wales, *Journal of Education Policy*, 15, 5, 559-573

Gorard, S. (2001) International comparisons of school effectiveness: a second component of the 'crisis account'?, *Comparative Education* , 37, 3, 279-296

Gorard, S. (2002) Fostering scepticism: the importance of warranting claims, *Evaluation and Research in Education*, 16, 3, 136-149

Gorard, S. (2005) Academies as the 'future of schooling': is this an evidence-based policy?, *Journal of Education Policy*, 20, 3, 369-377

Gorard, S. (2006a) Towards a judgement-based statistical analysis, *British Journal of Sociology of Education*, 27, 1, 67-80

Gorard, S. (2006b) *Using everyday numbers effectively in research*, London: Continuum

Gorard, S. (2006c) Value-added is of little value, *Journal of Educational Policy*, 21, 2, 233-241

Lunt, P. (2004) The significance of the significance test controversy: comments on 'size matters', *American Economic Review*, 33, 5, 559-564

Luyten, H. (2006) An empirical assessment of the absolute effect of schooling: regression-discontinuity applied to TIMSS-95, *Oxford Review of Education*, 32, 3, 397-429

Mansell, W. (2006a) Puzzle of new OFSTED ratings, *Times Educational Supplement*, 9/6/06, p.6

Mansell, W. (2006b) Shock of low score drives heads to resign, *Times Educational Supplement*, 9/6/06, p.6

Paterson, L. (1997) A commentary on methods currently being used in Scotland to evaluate schools statistically, pp. 298-312 in Watson, K., Modgil, C. and Modgil, S. (Eds.) *Educational dilemmas: debate and diversity*, London: Cassell

Pugh, G. and Mangan, J. (2003) What's in a trend? A comment on Gray, Goldstein and Thomas (2001), *British Educational Research Journal*, 29, 1, 77-82

Rumberger, R. and Palardy, G. (2005) Test scores, dropout rate, and transfer rates as alternative indicators of high school performance, *American Educational Research Journal*, 42, 1, 3-42

Schagen, I. (2006) The use of standardized residuals to derive value-added measures of school performance, *Educational Studies*, 32, 2, 119-132

Slater, J. (2006) Anger as Ofsted's 'raised bar' bites, *Times Educational Supplement*, 26/5/06, p.15

Tymms, P. and Dean, C. (2004) *Value-added in the primary school league tables: a report for the National Association of Head Teachers*, Durham: CEM Centre

Ziliak, S. and McCloskey, D. (2004) Significance redux, *American Economic Review*, 33, 5, 665-675