UNIVERSITY^{OF}BIRMINGHAM

Serious doubts about school effectiveness

Gorard, Stephen

[Link to publication on Research at Birmingham portal](#)

A case against school effectiveness

Stephen Gorard

The School of Education
The University of Birmingham
s.gorard@bham.ac.uk

Numbers are like people; torture them enough and they will tell you anything.

Keywords

School effectiveness, measurement error, PLASC, league tables

Abstract

This paper considers the model of school effectiveness (SE) currently dominant in research, policy and practice in England (although the concerns it raises are international). It shows, principally through consideration of initial and propagated error, that SE results cannot be relied upon. By considering the residual difference between the predicted and obtained score for all pupils in any phase of education, SE calculations leave the results to be disproportionately made up of relative error terms. Adding contextual information confuses but does not help this situation. Having shown and illustrated the sensitivity of SE to this propagation of initial errors, and therefore why it is unworkable, the paper considers some of the reasons why SE has become dominant, outlines the damage this dominant model causes, and begins to shape alternative ways of considering what schools do.

A case against school effectiveness

Numbers are like people; torture them enough and they will tell you anything.

Abstract

This paper considers the model of school effectiveness (SE) currently dominant in research, policy and practice in England (although the concerns it raises are international). It shows, principally through consideration of initial and propagated error, that SE results cannot be relied upon. By considering the residual difference between the predicted and obtained score for all pupils in any phase of education, SE calculations leave the results to be disproportionately made up of relative error terms. Adding contextual information confuses but does not help this situation. Having shown and illustrated the sensitivity of SE to this propagation of initial errors, and therefore why it is unworkable, the paper considers some of the reasons why SE has become dominant, outlines the damage this dominant model causes, and begins to shape alternative ways of considering what schools do.

The dominance of the school effectiveness model

There a number of valid possible reasons for wanting to be able to judge school performance. In most developed countries, the majority of schools are publicly funded, and so the custodians of public money want to assess how well that money is being used, for example. Policy-makers will be interested in how well this public service is working, and what the impact has been of any recent reforms. Parents and students might want to use a measure of school quality when making educational choices. Heads and teachers might want feedback on what is working well and what is in need of improvement at their own schools. There are also, of course, a number of differing ways of judging school performance. Schools could be evaluated in terms of financial efficiency, student attendance, student enjoyment of education, future student participation in education, student aspiration, preparation for citizenship and so on. Another perfectly proper indicator of school success can be based on student scores in assessments intended to discover how much or how well students have learnt what is taught in the school. What is interesting is how dominant this last version of school effectiveness has become over the last 50 years in the UK and elsewhere. This paper looks at the dominant approach to evaluating school performance, presenting fatal flaws in its logic, and so arguing that it is time to stop using this now traditional but limited view of what schools are for.

For any set of schools, if we rank them by their student scores in assessments of learning (the actual comparability and validity of such assessments is discussed in a later section), then we would tend to find that schools at the high and low ends differed in more than their student assessments. Schools in areas with more expensive housing (or more local income in the US), schools that select their student intake by ability, aptitude or even religion, and schools requiring parents to pay for their child's attendance, will be more prevalent among the high scores. Schools with high student mobility, in inner-cities, taking high proportions of children living in poverty or with

a different home language to the language of instruction, may be more prevalent among the low scores. This is well known, and means that raw-score indicators are not a fair test of school performance. Some early studies of school effectiveness famously found very little or no difference at all in the outcomes of schools once these kinds of student intake differences had been taken into account (Coleman et al. 1966). Such studies, using either or both of student prior attainment and student family background variables, have continued since then (Coleman et al. 1982, Gorard 2000a), and continue today (Lubienski and Lubienski 2006). The differences in student outcomes between individual schools, and types and sectors of schools, can be largely explained by the differences in their student intakes. The larger the sample, the better the study, and the more reliable the measures involved, the higher percentage of raw-score difference between schools that can be explained (Shipman 1997, Tymms 2003). Looked at in this way, it seems that which school a student attends makes little difference to their learning (as assessed by these means).

However, over the past 30 years a different series of studies have come to an almost opposite conclusion, based on pretty much the same evidence. Starting with Rutter et al. (1979) in the UK, and perhaps a little earlier in the US, school effectiveness researchers have accepted that much or most of the variation in school outcomes is due to school intake characteristics. But they have claimed that the residual variation (any difference in raw-scores unexplained by student intake) is, or can be, evidence of differential school effectiveness (e.g. Nuttall et al. 1989, Gray and Wilcox 1995, Kyriakides 2008). Like the first set of studies, these have tended to become more sophisticated and more technical over time. But the fundamental difference in view remains. Is the variation in school outcomes unexplained by student background just the messy stuff left over by the process of analysis? Or is it large enough, robust and invariant enough over time, to be accounted a school 'effect'? Can we promote, reward, and reprimand schools and teachers on this basis. Almost by default the answer to the second question has been assumed by most research users to be 'yes' (Barber and Mourshed 2007, Sanders 2000). There has been generally weak opposition to the dominant technical model of school effectiveness, perhaps stemming from inability to understand the technicalities (such as in Slee et al. 1998).[1]

Governments, such as that in the UK at time of writing, generally assume that there is a school effect. In England, DCSF (2007) rightly report that in comparing the performance of schools we must recognise that pupils have different starting points when arriving at any school, that schools have different proportions of pupils at any starting point, and that other external factors will affect the progress made by pupils.[2] They conclude from this that their Contextual Value Added analysis (CVA) 'gives a much fairer statistical measure of the effectiveness of a school, and provides a *solid* basis for comparisons' (p.2, emphasis added). On this basis, school inspection grades are partly pre-determined, schools are lauded or criticised, and league tables are created to assisted parental choice (see later section). How does this CVA work?

---

[1] I exclude some chapters from this statement of inability to comprehend, most especially the chapter by Brown (1998), which I urge everyone to read.

[2] The Department for Children, Schools and Families is responsible for the organisation of schools and childrens' services in England.

CVA is based on a value-added (VA) score for each pupil, calculated as the difference between their own outcome point score and the median outcome score for all pupils with the same prior (input) score. For example, in Key Stage 2 to Key Stage 4 CVA, the average points score at KS2 is calculated for all KS4 pupils in all maintained schools (and non-maintained special schools) in England.[3] The average is of the scores ('fine grades') for each pupil in three core subjects (English, maths, and science). Then the 'best 8' (capped GCSE equivalent) KS4 score is calculated for each pupil. These figures yield the median KS4 score for each KS2 score. The difference between the median and the actual KS4 score for each pupil is their individual VA score. This difference is adjusted for the individual pupil characteristics, including sex, special needs, ethnicity, eligibility for free school meals (FSM), first language, mobility, precise age, whether in care, and an areal measure of the proportion of households on low income (IDACI – an index of deprivation). The result is further adjusted for the school-level mean prior attainment of each pupil's school, where the results are at the extremes (threshold effects), and by a 'shrinkage factor' determined by the number of pupils in each school cohort.[4]

More formally and precisely, the KS4 prediction for any pupil in 2007 is given as:[5]

162.1
+0.3807 * (the squared school average KS2 score)
-5.944 * school average KS2 score
+1.396 * (KS2 English points - school average KS2 score)
-0.109 * (KS2 maths points - school average KS2 score)
-27.1 (if in care)
-59.51 * IDACI score
-34.37 (if School Action SEN)
-65.76 (if Action Plus or statement of SEN)
-73.55 (if joined after September of year 10)
-23.43 (if joined not in July/August/September of years 7-9)
+14.52 (if female)
-12.94 * (age within year, where 31st August is 0 and 1st September is 1)
+ *for English as an additional language pupils only* (-8.328 -0.1428*(school average KS2 score)^2 + 4.93 * school average KS2 score)
+ ethnicity coefficient, from a pre-defined table

---

[3] Key Stage 2 leads to statutory testing at the end of primary education, usually for pupils aged 11. Key Stage 4 leads to assessment at age 16, currently the legal age at which a pupil can leave school.

[4] These variables are used by DCSF for a number of reasons, including the fact that they are available at an individual level with reasonably complete data. Of course, other variables might be useful both at individual level, such as parents' occupation, and at school level, such as qualifications of teachers. Indeed, analyses for other purposes quite properly use different combinations of variables. However, the critique of CVA and school effectiveness presented here does not depend on the precise variables used. Occupation is harder to classify, and generally less complete as a field than eligibility for free school meals, for example. The omission of potentially important measures of individuals and schools can lead to another form of bias in the results, by making the variables that are included appear more important.

[5] Whereas both the ethnicity coefficient and the FSM/ethnicity interaction are 0 for White pupils, they are 29.190 and 20.4600 respectively for Black African pupils, for example.

+ *for FSM pupils only* (-22.9 + FSM/ethnicity interaction, from a pre-defined table)

+ 1.962 * cohort average KS2 score

- 4.815 * standard deviation of cohort average KS2 score

Equivalent models apply to CVA calculations for other stages of schooling, such as KS1 to KS2. Is the claim that a complex calculation such as this provides a *solid* basis for comparing school performance actually true?

Errors in the data

This kind of calculation looks very neat, if somewhat complex, and the logic seems plausible. SE models like CVA take the prior attainment and context of the student into account in order to judge their progress during one phase of schooling. This should be a better measure of the relative success than the raw-score results. Of course, the process depends heavily on the quality of the data used in the calculation. If the data is complete, correct, and an excellent measure of what it is intended to measure, then the process of calculating school effects in this way looks and sounds as though it has merit. Unfortunately, the kinds of datasets used for the job are necessarily incomplete, and contain both inaccuracies and errors in measurement. This section continues the example of CVA (above) as an illustration of the range and importance of these errors. The following section then shows how these errors propagate through the process of computation, making the results of school effectiveness calculations rather meaningless.

The first consideration is the completeness of the kinds of data needed for school effectiveness calculations. CVA in England is calculated using two linked official datasets − the National Pupil Database (NPD) and the Pupil Level Annual School Census (PLASC).[6] All schools are required by law to provide figures for these in the January of each school year, further data is added from existing official sources, and funding for the school hinges on their completion. PLASC contains a record for every pupil in maintained schools in England, detailing their background characteristics, including periods in-care, special needs status and first language. It also has some attainment data. NPD holds individual records on every pupil in maintained schools in England, detailing their examination and assessment entry and attainment, and also has some background data. They provide a wonderful and welcome resource for the researcher, at least the equal of equivalent datasets in other developed countries. Nevertheless, the records are not complete.

There are missing cases in the data, some by design such as those 7% of pupils attending private schools, and those educated at home. In addition, there will be small number of cases in transition between schools, or may otherwise not be in, or registered for, a school. Further, although both databases ostensibly contain records for all other pupils, in some years around 10% of the individual pupil records are un-matched across the two databases (see analysis by Gorard and See 2009, for example). This means, of course, that their background and attainment data cannot be

---

[6] For further explanation, contact the PLASC/NPD User Group (PLUG) at
http://www.bris.ac.uk/Depts/CMPO/PLUG/.

matched. The same thing happens trying to match cases across phases of schooling for the same pupils. In 2007 for example, the dataset for the Key Stage 4 (KS4 or 15-year-old) cohort contained records for 673,563 pupils, but nearly 10% of these could not be matched with the records of the same pupils from an earlier Key Stage such as KS2 (when they were 10-year-olds in the their final year of primary school). Any pupil moving to or from one of the other home countries of the UK such as Wales, where some statutory testing has been abolished, will have missing scores for one or more Key Stages. Any pupil moving from a private school, from a non-formal educational setting, or from outside the UK will similarly have no matching record of prior attainment at school on the PLASC/NPD system. In summary, perhaps nearly 10% of children will be missing from the databases completely, up to 10% will have a missing prior attainment record, and up to 10% will not a matched record in either PLASC or NPD. There will be some overlap between these missing cases, but this already represents a far from complete dataset.

The second consideration is the data missing even from those cases that do have records in the databases. In the 2007 PLASC/NPD datasets used to calculate CVA, every KS4 variable, including both the contextual and attainment variables, had a high proportion of missing cases. For example, whether a pupil was in-care had at least 80,278 values missing (12% of all cases). At least 75,944 were missing a code for free school meal eligibility (an important indicator of family poverty for CVA purposes). This represents over 11% of cases. Even when data does not appear as missing, it is effectively missing, such as the codes 'Refused' and 'Not obtained' which are additional to all data on pupil ethnic background which are coded as missing. There is again some overlap between these missing cases, but only some. For example, if we delete from the 2007 PLASC/NPD all cases missing data on FSM, in-care, special needs, sex and/or ethnicity data then the database drops in size to 577,115 pupils (or 85% of its apparent size, which was already itself incomplete as explained above).[7] If we consider *all* of the variables used in CVA, including further contextual variables such as pupil first language, and the attainment scores for each subject and grade (there are many of these), it is probable that less than 50% of the children of England in any age cohort have a record in all relevant databases that is complete in terms of all key variables.

One of the reasons for using area-based measures such as the index of deprivation (IDACI) is that they can replace missing data for individuals to some extent. However, this geographical approach suffers from two clear defects. First, it introduces a kind of ecological fallacy by assuming that everyone has the modal characteristics of the other people in the area where they live. Second, it relies on knowing the post code (area or ZIP code) of all individuals anyway. In the 2007 PLASCC/NPD, at least 69,902 (well over 10%) of the IDACI scores are missing because the address of the pupil is unknown. This then also introduces a clear error in at least one variable for *all* pupil records. The IDACI scores for all pupils, and as used in the CVA model, are calculated on the basis of scores for all households in England. Since the dataset used for this purpose does not, in fact, contain data for all households this means that all IDACI scores have an error component due to missing data over and above any errors in measuring household income (Gorard 2008a). Then

---

[7] Author's analysis using 2007 datasets for the purposes of this paper.

we need to realise that all of this missing data occurs not only in the KS4 datasets when the pupil is 15 or 16 but also in any other matched dataset such as KS2 used for the prior attainment scores when the pupil was aged 10 or 11. It is clear that missing data is a huge problem for any analysis of PLASC/NPD.

In practice, missing cases are simply ignored, and missing values are replaced with a default substitute – usually the mean score or modal category (and male for sex of pupil). So, the DCSF analysts assume that pupils without IDACI scores (usually because they have no post code) live in average income neighbourhoods, and that where we do not know when a pupil joined their present school we should assume that they have been in attendance for a long time. Anyone whose eligibility for FSM is not known is assumed not to be living in poverty, anyone without a KS2 or KS4 exam score is an average attainer, and so on. These kinds of assumptions have to be made in order not to lose the high number of cases with at least one missing value in a critical variable. But these are very questionable assumptions. There is plenty of evidence of differences between pupils with complete and incomplete values in such datasets (Amrein-Beardsley 2008). And making these unjustified assumptions then means that a very high proportion of cases are very likely to have an incorrect value in at least one critical variable.

How good then is the data that is not missing? Assessment via examination, project, coursework or teacher's grading is an imperfect process. There are huge and well-documented issues of comparability in assessment scores between years of assessment, curriculum subjects, modes of assessment, examining boards, and types of qualifications (among other issues, see Nuttall 1979, Newton 1997, Gorard 2000b). In fact, public assessment is generally handled well in England, and the kinds of high profile errors reported by Ofqual and others, such as up to 45% of candidates awarded the wrong grade in an extreme case, (Stewart 2009), are understandable in the light of a complex national testing and regulatory system (see below). To some extent these problems are coming to light because key figures at QCA decided that the public should be a given a more realistic picture of what test and exam 'standards' mean (http://www.ofqual.gov.uk/files/2009-03-18-national-curriculum-test-reviews-2000-7.pdf). Moderation will be imperfect and mistakes will be made. But we must assume a reasonable level of error in any assessment data of the kind used to calculate CVA.

Even when the system correctly assigns grades to pupils in their assessments we cannot be sure that they are free from error for a number of reasons. If we take the underlying competence of the pupil as the true measure wanted in an assessment, even a perfect assessment instrument could lead to error in the achieved measure due to differences in the setting for the assessment (a fire alarm going off in one examination hall, for example), time of day, inadvertent (and sometimes deliberate) teacher assistance, the health of the candidate, and so on. Competence is not an easy thing to measure, unlike the length of the exam hall or the number of people in it. However well-constructed the assessment system, we must assume a reasonable level of measurement error in the results.

Then the CVA analyst is faced with issues of aggregation and comparability. For example, the KS4 analysis involves GCSEs handled by different examining boards, sometimes taken via modules in different years, and for all different subjects and tiers of entry. Some GCSEs will be short courses, counting for half of credit of full GCSEs.

Even if an analyst is fairly sure about the comparability and reliability of such scores, these will have to be aggregated with results from an increasing number of different qualifications. In 2007, these included GNVQ Intermediate, NVQ, National Certificate in Business, BTEC, Key Skills, Basic Skills, and Asset Language Units. These all have to be converted to the common 'currency' of point scores, despite the fact that their grading structures are completely different. No one should try to claim that this aggregation to 'best 8' points scores does not add further errors to those catalogued so far. Issues of comparability are widely known, acknowledged and international in nature (Lamprianou 2009).

The same kind of consideration applies to any contextual variables. Even in NPD/PLASC with a simple binary code for sex, a few pupils are coded as male in one and female in the other database (more have nothing coded, and one or two have an invalid code, presumably from a data entry error). The error component in variables such as FSM, ethnicity, first language, and perhaps most particularly SEN, is even greater. Special educational needs, for example, are represented by a variable having three possible sources (School Action, Action Plus, or a statement). Some of these are the responsibility of the school, and some are sensitive to the actions of parents motivated to gain extra time in examinations for their children. The number of pupils with recorded SEN shows huge variation over years in the same schools, and appear in very different proportions in different parts of England (Gorard et al. 2003). Ethnic groups (based on 19 categories for CVA) are notoriously difficult to classify (Gorard 2008a). Here they are used in interaction with FSM eligibility (itself an incomplete measure). First language is almost as complex to classify as ethnic group. Is it home language, language of origin, or language of choice? Here it is used in interaction with prior attainment scores, since having a language other than English is calculated by the CVA model to be a disadvantage for low prior attainers, but not for high attainers. Where variables are used in interaction like this, to calculate CVA, an error in either one of them leads to an error in the combined result.

Once all of the relevant measurements have been achieved, they must be coded, entered into the databases and stored in binary floating-point format. Each step in this process allows the introduction of further errors. Coding data is subject to a low level of error even when conducted diligently, and not all such errors will be spotted by quality control systems dealing with hundreds of variables relating to millions of pupils every year. Then the data must be entered (transcribed) and low level errors are liable to creep in again. Data can be corrupted in storage (magnetic dropout undetected by parity checks and similar) and in sorting and matching of cases (most often caused by incorrect selection of rows or columns). Even a value for a pupil that is present and entered and stored 'correctly' is liable to be in error, due to the change in number base and the finite number of binary digits used to store it. The simple decimal fraction 0.1, for example, cannot be exactly represented in the binary numbering system used by computers and calculators. Thus, representing 0.1 in one byte (8 bits) would lead to an error of over 6% even in an otherwise perfect measurement. All numbers are generally stored in floating point form, involving a fractional mantissa and a binary exponent. Thus, the problem of representational errors can happen with any figure, whether it is an integer in denary or not. However many bits are allocated by a computer to storage of a number there will be, by definition, an infinite number of denary values that cannot be stored precisely.

Increased accuracy decreases these representational errors, but can not eliminate them.

At the end of all this it is hard to believe that any pupil record will be free from all errors, with so many areas for errors to creep into the data from missing cases to conversion to binary. However, the CVA formula used by DCSF uses the measurements in calculations supposed to be accurate to at least four decimal places. It multiplies individual point scores coefficients represented to decimal places (i.e. claiming to be correct to 5/1000ths of a point) and multiplies them by coefficients with four decimal places (such as +0.3807). So in the example on the CVA website, the first term after the constant in the CVA formula could be 29.56 squared times 0.3807. This would be 332.6532235 (correct to 5 parts in 10 million). This is pseudo-quantification of the worst kind. There is no way that the initial figures are accurate enough to sustain this kind of calculation, as the next section illustrates. CVA in England has been used as an illustration of the problems in the data even for an excellent dataset. Similar problems or worse appear in other official datasets in the UK (Gorard 2008a) and in other countries like the USA (Sanders and Horn 1998, p.248).

It is worth pointing out at this stage in the argument that any analysis using real data with some combination of the inevitable measurement errors described so far will lead to an incorrect result. Of course, the more accurate the measures are the closer to the ideal correct answer we can be. However, we have no reason to believe that any or all of these sources of error lead to random measurement error (of the kind that might come from random sampling variation, for example). Those refusing to take part in a survey, those not registered at school, those unwilling to reveal their family income or benefit (for free school meal eligibility purposes) cannot be imagined as some kind of random sub-set of the school population. Like every stage in the error generation process described so far, they are not random in nature, occurrence or source. What happens to these errors in a school effectiveness calculation?

The propagation of errors

For any real measurement that we use for analysis we must assume the possibility of measurement error. Measurement error in this context means a difference between the ideal or perfectly isomorphic representation of something and our achieved measure. If someone actually has three children but our measurement claims that they have two children, then our measurement of the number of children is in error by one. This simple discrepancy is often termed the absolute error. A more useful way of envisaging such an error is as a fraction of the measurement itself – the relative error. In this example, the relative error is 1/2. In trying to measure 3 we achieve a measure of 2 which is out by 1. If we were out by 1 in attempting to measure the number of children in the entire country, this would not be such a serious measurement error, and the relative error would be much smaller than 1/2.[8]

---

[8] Strictly, the relative error is 1/3 based on the true value we are trying to measure. In practice of course we do not know this value else there would be no error, so all relative errors are here based on the achieved measure instead.

As a consequence of the errors discussed in the previous section, imagine for the sake of argument that all measures such as pupil prior attainment used in CVA were only 90% accurate, having a relative error of 1/10. What would this mean? In itself, it tells us what we already know – that the score for any pupil cannot be guaranteed to be accurate. We should not treat a score for one pupil of 70 as being substantially difference in practice from a score of 73 for another pupil. The difference between them is smaller than the error bound of each. On the other hand, it means that a score of 70 *can* be treated as substantially different from a score of 100, since the difference is greater than the error bound. Put another way an achieved score of 70 in the database could be between 63 and 77 in reality (+/-10%). An achieved score of 100 could be between 90 and 110 in reality. Since 90 is still larger than 77 we can proceed with some confidence that the score represented by 100 really is larger than the score of 70. For normal descriptive purposes in education and social science a relative error of 10% in our achieved figures is acceptable. But what happens when we feed scores such as these into a school effectiveness calculation like CVA?

Errors are said to 'propagate' through calculations, meaning that everything we do with our achieved measures we also do with their measurement errors. The relative error changes as a consequence. If we have two numbers X and Y measured imperfectly as x and y with corresponding absolute errors ε_x and ε_y then:

$$x = X \pm \varepsilon_x$$

and

$$y = Y \pm \varepsilon_y$$

When we attempt to calculate X-Y, we actually get x-y which is equal to $(X \pm \varepsilon_x) - (Y \pm \varepsilon_y)$. The upper bound for this is $X-Y + \varepsilon_x + \varepsilon_y$. Put another way, since we do not know whether the errors in either number are positive or negative when we subtract we may be adding the error components (and vice versa of course).

In England, the model for contextualised value-added analysis used by the Department for Children, Schools and Families (DCSF) involves finding for all pupils 'the difference (positive or negative) between their predicted and actual attainment' (DCSF 2007, p.7). The predicted attainment for any one pupil is based on the average gain score for all pupils with the same prior attainment (adjusted for contextual information). The difference between any pupil's predicted and actual attainment will tend to be insubstantial for two reasons. First, the predicted and actual attainment scores are not just of same magnitude, and using the same points system. They are designed to be as close as possible to each other. Second, if the predicted and actual attainment scores were not very close for a majority of pupils then the model would not be any good. This means that the figure computed for the pupil value-added score is usually very small, perhaps even negligible, in comparison to the attainment scores from which it is calculated. CVA subtracts the predicted and actual attainment to create a much smaller figure, but adds their maximum errors (since we do not know if the errors are positive or negative).

For an illustration of the importance of this propagation of errors, imagine a pupil with an actual points score of 100 for attainment at KS4, but with a predicted points

score of 99. The prediction is a good one in that it is close, but the pupil appears to have made marginally more progress than expected. Both scores are assumed to be 90% accurate (see above). This relative error of only 10% is a very conservative estimate given the multiple sources of error described in the previous section, and the scale of missing data. The predicted score, based on all of the CVA variables in isolation and in interaction will have a much larger error component than this in reality. But even an error of 10% means that the actual score for this pupil could be anywhere between 90 and 110 in reality, and the predicted score ought to be anywhere from 89.1 to 108.9. This means that the real residual score for this pupil (their CVA score) could be anything from +20.9 (110-89.1) to -18.9 (90-108.9). The maximum relative error in the calculated answer of +1 is a massive 3,980%. By subtracting two similar numbers with an acceptable level of initial error (10%) we are left with an 'answer' composed almost entirely of error (3,980%). We genuinely have no idea whether this pupil has done better or worse than expected. There is no way that such a result could be used for any practical purpose. If the initial relative error in either the actual or the predicted score is greater than 10%, as it almost certainly would be in reality, the error in the CVA result would be even greater than this 40 times the size of the result itself.

Where the actual and predicted score are the same for any pupil (i.e. when the CVA model works well), the residual score is zero and so the relative error in the result is infinite. As the achieved and predicted scores diverge the relative error in the residual tends to decline. But this then means that the CVA model, which is meant to make *accurate* predictions, is not working well. If the predictions are so far out that we can begin to ignore the error components is this better or worse for the school effectiveness model? In order to retain something like the relative error of 10% in the original scores, the CVA prediction would have to be out by a long way from the achieved result. For example, a predicted score of 50 with a 10% initial error represents a range of 45 to 55. An actual score for the same pupil of 100 with a 10% initial error represents a range of 90 to 110. This means that the real residual score for this pupil (their CVA score) could be anything from +65 (110-45) to +35 (90-55). This yields a maximum relative error of 60% in the resulting CVA score of +50. So even when the CVA prediction is way out, as in this example, an initial error of 10% propagates to 60% via simple subtraction. If we assume that the school effectiveness model is capturing anything sensible at all, this pupil can be deemed to have done very well (or to have done very badly in the prior assessment, or both). This is true even if the maximum error applies. How can we tell whether any CVA score (for pupil, teacher, department, school or area) is of this kind, where we cannot be sure about the precise figure but we can be sure that the result is so far away from that predicted as to dwarf any error component?

The allure of technical solutions

Unfortunately the field of school effectiveness research works on the invalid assumption that errors in the data are random in nature and so can be estimated, and weighted for, by techniques based on random sampling theory. These techniques are fatally flawed, in their own terms, even when used 'correctly' with random samples (Gorard 2010). The conditional probabilities generated by sampling theory tell us, under strict conditions and assumptions, how often random samples would generate a

result as extreme or more extreme as the one we might be considering. The p-value in a significance test tells analysts the probability of observing a score as or more extreme as the measure they achieved, assuming that the score is actually no different from zero (and so that the divergence from zero is the result of random sampling variation alone). Of course, this conditional probability of the data given the nil null hypothesis is not what the analysts want. In a school effectiveness context such as the ones outlined above, the analyst wants to know whether the CVA score (the residual, whether for individual or school) is large enough to take note of (to dwarf its relative error). They actually want the probability of the null hypothesis given the data they observed. They could convert the former to the latter using Bayes' Theorem, as long as they already knew the underlying and unconditional probability of the null hypothesis anyway. But they cannot know the latter. So they imagine that the probability of the data given the null hypothesis is the same as, or closely related to, the probability of the null hypothesis given the data. They then use the p-value from significance tests to 'reject' the null hypothesis on which the p-value is predicated. This *modus tollens* kind of argument does not work with likelihoods, for a number of reasons, including Jeffrey's so-called paradox that a low probability for the data can be associated with a high probability for the null hypothesis, or a low one, or a mid-range value, and *vice versa*. It depends on the underlying probability of the null hypothesis - which we do not know.

So, even used as intended, p-values can not help most analysts in the SE field. The same applies to standard errors, and confidence intervals and their variants. But the situation is worse than this because in the field of school effectiveness, these statistical techniques based on sampling theory are hardly ever used *as intended*. Most commonly, the sampling techniques are used with population figures such as NPD/PLASC. In this context, the techniques mean nothing.[9] There is no sampling variation to estimate when working with population data (whether for a nation, region, education authority, school, year, class, or social group). There are missing cases and values, and there is measurement error. But these are not generated by random sampling, and so sampling theory cannot estimate them, adjust for them, or help us decide how substantial they are in relation to our manifest data.[10]

---

[9] This does not prevent the widespread abuse of random sampling techniques with population data. A couple of recent examples will have to suffice. Hammond and Yeshanew (2007, p.102) base their analysis on a national dataset, but say 'Although no actual samples have been drawn… Statistical checks were carried out and no significant difference between the groups was found'. They then present a table of standard errors for this population data (p.102). They have learnt to use multi-level modelling but clearly forgotten what significance means and what a standard error is. Similarly, Thomas, Peng and Gray (2007) examined data from one school district in England (and so a population, in statistical terms). Yet they report (p.271) that 'the pupil intake and time trend explanatory variables included in the fixed part of the value-added model (Model A) were statistically significant (at 0.05 level)'.

[10] This misuse of sampling theory with population data has sometimes been defended by saying that the population figures are somehow a random sample of a theoretical 'super-population'. In the example of PLASC/NPD, then, the school pupils are imagined as a random sample of all the children that could have been born to their parents, and the analyst seeks to generalise the findings to those unborn and never born children. But why should the born children be a *random* subset of the unborn? What does that even mean in real-life? Do politicians and parents know that this is what such statisticians mean? And why would anyone want to generalise to a non-existent and never-to-be-born group anyway? This is an example of the lengths than some analysts will go to in defending their sampling theory 'toys'. The approach has long been discredited (Camilli 1996, Gorard 2008b).

Despite all this, DCSF use and attempt to defend the use of confidence intervals with their population CVA data. A confidence interval, remember, is an estimate of the range of values that would be generated by repeated random sampling, assuming fro calculations purposes that our manifest score is the correct one. It has no relevance at all to population data like PLASC/NPD. It is of no real use to an analyst, even when calculated with for a random sample, for the same reasons as for p-values.[11] The analyst wants a probable range for the true value of the estimate, but to get this they would have to have access to underlying data that is never available to them. And as with p-values, it does not even make sense to calculate a confidence interval for population data of any kind. Confidence intervals are therefore of no use in standard school effectiveness research.[12]

However, the field as a whole simply ignores these quite elementary logical problems, while devising more and more complex models comprehended by fewer and fewer people. Perhaps the most common inappropriate complex technique used in this field is multi-level (hierarchical linear) modelling. This technique was devised as one of many equivalent ways of overcoming the correlation between cases in cluster-randomised samples (Gorard 2009a). This, like all other techniques based on sampling theory, is of no consequence for school effectiveness work based on population figures. Advocates now claim that such models have other purposes – such as allowing analysts to partition variation in scores between levels such as individuals, schools and districts. But such partitioning can, like overcoming the inter-correlation in clusters, be done in other and generally simpler ways. Anyway, the technique is still pointless. Most such models do not use districts or areas as a level, and those that do tend to find little or no variation there once other levels have been accounted for (Smith and Street 2006, Tymms et al. 2008). We know that pupil-level variables, including prior attainment and contextual values, are key in driving school outcomes. The question remains, therefore, whether there is a school effect. If our pupil-level predictions of subsequent attainment are less than perfect, we could attribute much of the residual unexplained variation to the initial and propagated measurement error in our data. To use multi-level modelling to allocate most of this residual variation to a 'school effect' instead is to assume from the outset that which the modelling is supposed to be seeking or testing.

So why does school effectiveness seem to work?

---

[11] Some commentators and even some purported training resources suggest that a confidence interval is a band within which we can be reasonably confident the true population figure appears. This is a simple error of understanding. All confidence intervals have the manifest score at their centre, and are clearly a band of likely scores we would achieve if the random sampling that led to the manifest score were repeated, and using the manifest score is our best (only) guess so far. We have no idea where the true population figure actually is, other than from that guess, unless we have the population figures (as we do in this paper). If we have the population figures (as we do in this paper) we do not need confidence intervals, and they make no sense then anyway.

[12] Some purported authorities on school effectiveness still erroneously propose the use of confidence intervals with school effectiveness scores based on population figures (e.g. Goldstein 2008).

Why, if the foregoing is true, do so many analysts, policy-makers, users and practitioners seem to believe that school effectiveness yields useful and practical information? It is tempting to say that perhaps many of them have not really thought about the process and have simply bought into what appears to be a scientific and technical solution to judging school performance. I use the term 'bought' advisedly here because part of the answer might also lie in the money to be made. In England, school effectiveness has become an industry, employing civil servants at DCSF and elsewhere, producing incentives for teachers deemed CVA experts in schools, creating companies and consultants to provide data analysis, paying royalties to software authors, and funding for academics from the taxpayer. A cynical view would be that most people in England do not understand CVA, and a high proportion of those that do stand to gain from its use in some way.

There is sometimes no consistent adherence to school effectiveness as a model, even among individual policy-makers and departments. Some of the schools required by DCSF in 2008 to take part in the National Challenge, because their (raw-score) results were so poor, were also sent a letter from DCSF congratulating them on their high value-added results and asking them to act as models or mentors for emergent Academies. The 'paradox of the National Challenge Scheme' continues (Maddern 2008, p.23). Again, a cynic might say that users use raw scores when it suits them (traditional fee-paying schools seem uninterested in value-added while often having very high raw scores, for example), and they use value-added when that paints a better picture.

However, it is possible that the problem stems chiefly from our lack of ability to calibrate the results of school effectiveness models against anything except themselves. In everyday measurements of time, length, temperature and so on we get a sense of the accuracy of our measuring scales by comparing measurements with the qualities being measured (Gorard 2009b). There is no equivalent for CVA (what Amrein-Beardsley refers to as criterion-related validity). The scores are just like magic figures emerging from a long-winded and quasi-rational calculation. Their advocates claim that these figures represent 'solid' and fair school performance measures, but they can provide nothing except the purported plausibility of the calculation to justify that. Supposing, for the sake of argument, that the calculation did not work for the reasons given in this paper so far. What would we expect to emerge from it? The fact that the data is riddled with initial errors and that these propagate through the calculation does not mean that we should expect the results for all schools to be the same, once contextualised prior attainment is accounted for. The bigger the deviations between predicted and attained results, of the kind that SE researchers claim as evidence of effectiveness, the more this could also be evidence of the error component. In this situation, the bigger the error in the results the bigger the effect might appear to be to some. So, we cannot improve our approach to get a bigger effect to outscore the error component. Whatever the residuals are we simply do not know if they are error or effect. We do know, however, that increasing the quality and scale of the data is associated with a decrease in the apparent school effect (Tymms 2003).

If the VA residuals were actually only error, how would the results behave? We would expect CVA results to be volatile and inconsistent over years and between key stages in the same schools. This is what we generally find (Hoyle and Robinson 2003, Tymms and Dean 2004, Kelly and Monczunski 2007). Of course, in any group of

schools under consideration, some schools will have apparently consistent positive or negative CVA over a period of time. This, in itself, means nothing. Again imagine what we would expect if the 'effect' were actually all propagated error. Since CVA is zero-sum by design, around half of all schools in any one year would have positive scores and half negative. If the CVA were truly meaningless, then we might expect around one quarter of all schools to have successive positive CVA scores over two years (and one quarter negative). Again, this is what we find. *Post hoc*, we cannot use a run of similar scores to suggest consistency without consideration of what we would expect if the scores meant nothing. Thomas, Peng and Gray (2007) looked at successive years of positive VA in one England district from 1993-2002. They seemed perplexed that 'it appears that only one in 16 schools managed to improve continuously for more than four years at some point over the decade in terms of value-added' (p.261). Yet 1 in 16 schools with four successive positive scores is exactly how many would be predicted assuming that the scores mean nothing at all (since 2^{-4} equals 1/16).

Leckie and Goldstein (2009) explain that VA scores for the same schools do not correlate highly over time. A number of studies have found VA correlations of around 0.5 and 0.6 over two to five years for the same schools. Whatever it is that is producing VA measures for schools, it is ephemeral. A correlation of 0.5 after two years means that only 25% of the variation in VA is common to those years. Is this really any more than we would expect by chance? What is particularly interesting about this variability is that it does not appear in the raw scores. Raw scores for any school tend to be very similar from year to year, but the 'underlying' VA is not. Is this then evidence, as Leckie and Goldstein (2009) would have it, that VA really changes that much and so quickly, or does it just illustrate again the central point in this paper that VA is very sensitive to the propagation of relative error?

The coefficients in the CVA model, fitted *post hoc* via multi-level regression, mean nothing in themselves. Even a table of complete random numbers can generate regression results as coherent (and convincing to some) as SE models (Gorard 2008b). With enough variables, combinations of variables and categories within variables (remember the 19 ethnic groups in interaction with FSM in CVA, for example) it is possible to create a perfect multiple correlation (R^2=1.00) from completely nonsensical data. In this context, it is intriguing to note the observation by Glass (2004) that one school directly on a county line was attributed to both counties in the Tennessee Value Added Assessment System and two VA measures were calculated. The two results were completely different – suggesting perhaps that they did not really mean anything at all. Even advocates and pioneers of school effectiveness admit that the data and models we have do not allow us to differentiate, in reality, between school performances. 'Importantly, when we account for prediction uncertainty, the comparison of schools becomes so imprecise that, at best, only a handful of schools can be significantly separated from the national average, or separated from any other school' (Leckie and Goldstein 2009, p.16).

Of course, the key calculation underlying CVA is the creation of the residual between actual and predicted pupil scores. Since this is based on two raw scores (the prior and current attainment of each pupil), it should not be surprising to discover that VA results are highly correlated with each of these raw scores (Gorard 2006, 2008c). The scale of this correlation is now routinely disguised by the contextual figures used in

CVA, but it is still there. In fact, the correlation between prior and current attainment is the same size as the correlation between prior attainment and VA scores. Put more simply, VA calculations are flawed from the outset by not being independent enough of the raw scores from which they are generated. They are no more a fair test of school performance than raw scores are.

Damage caused by school effectiveness

Does any of this matter? I would argue that it does. Schools, heads and teachers are being routinely rewarded or punished on the basis of this kind of evidence. Teachers are spending their time looking at things like departmental value-added figures and distorting their attention to focus on particular areas or types of pupils. School effectiveness results have been used to determine funding allocations and to threaten schools with closure (Bald 2006, Mansell 2006). The national school inspection system in England, run by OFSTED, starts with a CVA, and the results of that analysis partly pre-determine the results of the inspection (Gorard 2008c). Schools are paying public funds to external bodies for value-added analyses and breakdowns of their effectiveness data. Parents and pupils are being encouraged to use school effectiveness evidence (in league tables, for example) to judge their schools and potential schools. If, as I would argue, the results are largely spurious this means a lot of time and money is wasted and, more importantly, pupils' education is being needlessly damaged.

However, the dangers of school effectiveness are even greater than this. School effectiveness is associated with a narrow understanding of what education is for. It encourages, unwittingly, an emphasis on assessment and test scores – and teaching to the test – because over time we tend to get the system we measure for and so privilege. Further, rather than opening information about schools to a wider public, the complexity of CVA and similar models excludes and so disempowers most people. These are the people who pay tax for, work in, or send their children to schools. Even academics are largely excluded from understanding and so criticising school effectiveness work (Normand 2008). Relevant academic work is often peer-reviewed and 'quality' checked by a relatively small clique. School effectiveness then tends to monopolise political expertise on schools and public discussion of education, even though most policy-makers, official bodies like OFSTED, and the public simply have to take the results on trust.

The widespread use of CVA for league tables, official DCSF performance data, and in models of school effectiveness also has the inadvertent impact of making it harder to examine how well schools are doing with different groups of pupils. One of the main reasons for initially setting up a free (taxpayer-funded), universal and compulsory system of schools was to try and minimise the influence of pupil family background. The achievement gaps between rich and poor, or between ethnic and language groups, give schools and society some idea of how well that equitable objective is being met. What CVA does is to recognise that these gaps exist but then makes them invisible by factoring them into the VA prediction. It no longer makes sense to ask whether the CVA is any different in a school or a school system for rich and poor, or different ethnic and language groups. DCSF (2007) appear to recognise this danger when they say (in bold, p.2) 'CVA should not be used to set lower expectations for any pupil or

group of pupils'. This means – bizarrely - that a school with a high level of poverty will be correctly predicted to have equivalently lower outcomes, but at the same time must not 'expect' lower outcomes.

Finally, for the present section, it is important to recall that VA, CVA and the rest are all zero-sum calculations. The CVA for a pupil, teacher, department, school, or district is calculated relative to all others. Thus, around half of all non-zero scores will be positive and half negative. Whether intentionally or not, this creates a system clearly based on competition. A school could improve its results and still have negative CVA if everyone else improved as well. A school could even improve its results and get a worse CVA than before. The whole system could improve and half of the schools would still get negative CVA. Or all schools could get worse and half would still get positive CVA scores. And so on. It is not enough to do well. Others have to fail for any school to obtain a positive result. Or more accurately, it is not even necessary to do well at all; it is only necessary to do not as badly as others. This is a ridiculous way of calculating school performance *sui generis*, as shown in this paper so far. But why, in particular, design the monitoring system like that at the same time as asking schools in England to form partnerships and federations, and to co-operate more and more in the delivery of KS3 and the 14-19 Reform Programme?

What does it all mean?

The whole school effectiveness model, as currently imagined, should be abandoned. It clearly does not and could not work as intended, so it causes all of the damage and danger described above for no good reason. It continues partly as a kind of Voodoo science (Park 2000), wherein adherents prefer to claim they are dealing with random events, making it easier to explain away the uncertainty and unpredictability of their results. But it also continues for the same reasons as it was created in the first place. We want to be able to measure school performance, and we know that mere raw-score figures tell us largely about the school intake. However, we must not continue with school effectiveness once aware of its flaws simply because we cannot imagine what to do instead. The purpose of this paper is to try and end the domination of school policy and much of research by the standard school effectiveness model. That is a big task for one paper. It is not my intention here to provide a fully worked-out alternative.[13]

We perhaps need to re-think what we mean by a school effect. In traditional models, a school effect refers to the difference going to one school makes in comparison to going to another school. There are many other possible meanings we could operationalise, including what difference it makes going to one school as opposed to not going school at all. We need to decide whether we are happy for a school effect to be zero-sum, whether a school is really a proper unit of analysis, and how we will estimate the maximum propagation of errors. I would like to see much greater care in

---

[13] One promising avenue is based on regression discontinuity (e.g. Luyten 2006). This has the major advantage over CVA of not being zero-sum in nature. All schools could improve and be recognised for this (and vice versa), and groups of schools or whole districts can be assessed as co-operative units in which the success of any unit adds to the success of any other. Surely something like this is better for now and for the future.

the design of research and the collection of data, rather than effort expended on creating increasingly complex and unrealistic methods to analyse the poorer data from existing poor designs and models. We need more active designs for research, such as randomised controlled trials to find out what really works in school improvement, rather than *post hoc* data dredging. We need more mixed methods studies, and more care and humility in our proposed research claims. Education research is, rightly, a publicly-funded enterprise with potential impacts for all citizens. If our research is as poorly crafted as school effectiveness seems to be then we face two possible dangers. If it has no real-life impact then the research funding has been wasted. There is a significant opportunity cost. Worse, if it has the kind of widespread impact that school effectiveness has had for thirty years, then in addition to the waste of money, incorrect policy and practice decisions will be made, and pupils and families will suffer the consequences.

One clear finding that is now largely unremarked by academics and unused by policy-makers is that pupil prior attainment and background explain the vast majority of variation in school outcomes. This finding is clear because its scale and consistency over time and place dwarfs the error component in the calculation (largely because the error does not have a chance to propagate in the same way as for CVA analysis). Why is this not more clearly understood and disseminated by politicians? In England, we have built a system of maintained schools that remains loosely comprehensive, and is funded quite equitably (more so than the USA, for example), on a per-pupil basis adjusted for special circumstances. The curriculum is largely similar (the National Curriculum) for ages 5 to 14 at least, taught by nationally-recognised teachers with Qualified Teacher Status, inspected by a national system (OFSTED), and assessed by standardised tests up to Key Stage 3. Education is compulsory for all, and free at the point of delivery. In a very real sense it sounds as though it would not matter much which school a pupil attends, in terms of qualifications as an outcome. And indeed, that is what decades of research have shown is true.

Are parents and pupils being misled into thinking that which school they use does make a substantial difference? Perhaps, or perhaps qualifications at age 16 are not what parents and pupils are looking for when they think of a new school for a child aged 4 or even 10.[14] School choice research suggests that what families are really looking for is safety and happiness for their children (Gorard 1997). When thinking about moving a 10-year-old from a small primary school in which they are the oldest to a much larger, more distant secondary school with students up to the age of 19, security is often the major concern. This is why proximity can be seen as a rational choice. It is also possible that parents know perfectly well that raw-scores are not an indication of the quality of the school attended but of the other pupils attending. Using raw scores might be a rational way for a lay person to identify a school in which learning was an important part of everyday school life. Raw scores, like bus stop behaviour, are used as a proxy indication of school intake.

If so, several conclusions might follow. Politicians could disseminate the truth that in terms of traditional school outcomes it makes little difference which school a pupil attends. This might reduce the allure of specialisms, selection by aptitude or

---

[14] Or perhaps parents are smarter than policy-makers, realising that current VA scores for any school or phase are historical and tell them only what might have happened if their child had started at that school five years ago.

attainment, faith-based schools, and other needlessly divisive elements for a national school system. It could reduce the so-called premium on housing near to what are currently considered good schools, and reduce the journey times to schools (since the nearest school would be as good as the furthest). All of this would be associated with a decline in socio-economic and educational segregation between schools. SES segregation between schools has been a rising problem in England since 1997 (Gorard 2009b). Reduced segregation by attainment and by student background has many advantages both for schools and for wider society, as well as becoming a repeating cycle, making schools genuinely comprehensive in intake as well as structure, so giving families even less reason to look beyond their nearest schools. It would also mean that, on current figures, no schools would be part of the National Challenge. Schools are earmarked for the National Challenge if their KS4 raw score benchmark of pupils attaining the equivalent of five good GCSEs is less than 30%. Since the overall national figure is considerably higher than 30%, the National Challenge is less an indication of poor schools and more an indictment of the levels of academic segregation in the system. Redistributing school intakes solves the problem at a stroke.

Perhaps even more importantly, once policy-makers understand how CVA works and that they cannot legitimately use it to differentiate school performance, they may begin to question the dominance of the school effectiveness model more generally. We might see a resurgence of political and research interest in school processes and outcomes other than pencil-and-paper test results. Schools are mini-societies in which, according to surveys, pupils may learn how to interact, what to expect from wider society, and how to judge fairness (Gorard and Smith 2009). Schools seem to be a key influence on pupils' desire to take part in future learning opportunities (Gorard et al. 2007), and on their occupational aspirations (Gorard and Rees 2002). All of these outcomes have been largely ignored in three decades of school effectiveness research. It is time to move on.

References

Amrein-Beardsley, A. (2008) Methodological concerns about the education value-added assessment system, *Educational Researcher*, 37, 2, 65-75
Bald, J. (2006) Inspection is now just a numbers game, *Times Educational Supplement*, 26/5/06, p.21
Barber, M. and Mourshed, M. (2007) *How the world's best-performing school systems come out on top*, McKinsey & Co.
Brown, M. (1998) The tyranny of the international horse race, pp.33-47 in Slee, R., Weiner, G. and Tomlinson, S. (Eds.) *School Effectiveness for Whom? Challenges to the school effectiveness and school improvement movements*, London: Falmer Press
Camilli, G. (1996) Standard errors in educational assessment: a policy analysis perspective, *Education Policy Analysis Archives*, 4, 4
Coleman, J., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfield, F. and York, R. (1966) *Equality of educational opportunity*, Washington: US Government Printing Office
Coleman, J., Hoffer, T. and Kilgore, S. (1982) Cognitive outcomes in public and private schools, *Sociology of Education*, 55, 2/3, 65-76

DCSF (2007) *A technical guide to the contextual value added 2007 model*, http://www.dcsf.gov.uk/performancetables/primary_07/2007GuidetoCVA.pdf, accessed 16/12/08

Glass, G. (2004) *Teacher evaluation: Policy brief*, Tempe, Arizona: Education Policy Research Unit

Goldstein, H. (2008) Evidence and education policy – some reflections and allegations, *Cambridge Journal of Education*, 38, 3, 393-400

Gorard, S. (1997) *School Choice in an Established Market*, Aldershot: Ashgate

Gorard, S. (2000a) '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. (2000b) *Education and Social Justice*, Cardiff: University of Wales Press

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

Gorard, S. (2008a) Who is missing from higher education?, *Cambridge Journal of Education*, 38, 3, 421-437

Gorard, S. (2008b) *Quantitative research in education*, London: Sage

Gorard, S. (2008c) The value-added of primary schools: what is it really measuring?, *Educational Review,* 60, 2, 179-185

Gorard, S. (2009a) Misunderstanding and misrepresentation: a reply to Schagen and Hutchison, *International Journal of Research and Method in Education*, 32, 1

Gorard, S. (2009b) Measuring is more than assigning numbers, in Walford, G., Tucker, E, and Viswanathan, M. (Eds.) *Handbook of Measurement*, Sage, (submitted)

Gorard, S. (2009c) Does the index of segregation matter? The composition of secondary schools in England since 1996, *British Educational Research Journal*, (forthcoming)

Gorard, S. (2010) All evidence is equal: the flaw in statistical reasoning, *Oxford Review of Education*, (forthcoming)

Gorard, S. and Rees, G. (2002) *Creating a learning society?*, Bristol: Policy Press

Gorard, S. and See, BH. (2009) The impact of SES on participation and attainment in science, *Studies in Science Education*, 45, 1, 93-129

Gorard, S. and Smith, E. (2009) The impact of school experiences on students' sense of justice: an international study of student voice, *Orbis Scholae*, 2, 2, 87-105

Gorard, S., Taylor, C. and Fitz, J. (2003) *Schools, Markets and Choice Policies,* London: RoutledgeFalmer

Gorard, S., with Adnett, N., May, H., Slack, K., Smith, E. and Thomas, L. (2007) *Overcoming barriers to HE*, Stoke-on-Trent: Trentham Books

Gray, J. and Wilcox, B. (1995) *'Good school, bad school' Evaluating performance and encouraging improvement*, Buckingham: Open University Press

Hammond, P. and Yeshanew, T. (2007) The impact of feedback on school performance, *Educational Studies*, 33, 2, 99-113

Hoyle, R. and Robinson, J. (2003) League tables and school effectiveness: a mathematical model, *Proceedings of the Royal Society of London B*, 270, 113-199

Kelly, S. and Monczunski, L. (2007) Overcoming the volatility in school-level gain scores: a new approach to identifying value-added with cross-sectional data, *Educational Researcher*, 36, 5, 279-287

Kyriakides, L. (2008) Testing the validity of the comprehensive model of educational effectiveness: a step towards the development of a dynamic model of effectiveness, *School Effectiveness and School Improvement*, 19, 4, 429–446

Lamprianou, I. (2009) Comparability of examination standards between subjects: an international perspective, Oxford Review of Education, 35, 2, 205-226

Leckie, G. and Goldstein, H. (2009) *The limitations of using school league tables to inform school choice*, Working Paper 09/208, Bristol: Centre for Market and Public Organisation

Lubienski, S. and Lubienski, C. (2006) School sector and academic achievement@ a multi-level analysis of NAEP Mathematics data, *American Educational Research Journal*, 43, 4, 651-698

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

Maddern, K. (2009) Adding value, but still a challenge, *Times Educational Supplement*, 16/1/09, p.23

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

Newton, P. (1997) Measuring comparability of standards across subjects: why our statistical techniques do not make the grade, *British Educational Research Journal*, 23, 4, 433-449

Normand, R. (2008) School effectiveness or the horizon of the world as a laboratory, *British Journal of Sociology of Education*, 29, 6, 665-676

Nuttall, D. (1979) The myth of comparability, *Journal of the National Association of Inspectors and Advisers*, 11, 16-18

Nuttall, D., Goldstein, H., Presser, R. and Rasbash, H. (1989) Differential school effectiveness, *International Journal of Educational Research*, 13, 7, 769-776

Park (2000) *Voodoo science*, Oxford, OUP

Rutter, M., Maughan, B., Mortimore, P. & Ouston, J. (1979). *Fifteen thousand hours: Secondary schools and their effects on children*. London: Open Books.

Sanders, W. (2000) Value-added assessment from student achievement data, , *Journal of Personnel Evaluation in Education*, 14, 4, 329-339

Sanders, W. and Horn. S. (1998) Research findings from the Tennessee Value-added assessment system (TVAAS) database, *Journal of Personnel Evaluation in Education*, 12, 3, 247-256

Shipman, M. (1997) *The limitations of social research*, Harlow: Longman

Slee, R., Weiner, G. and Tomlinson, S. (1998) *School Effectiveness for Whom? Challenges to the school effectiveness and school improvement movements*, London: Falmer Press

Smith, P. and Street, A. (2006) *Analysis of secondary school efficiency: final report*, DfES Research Report 788

Stewart, W. (2009) 'Appalling' failures of English test markers, *Times Educational Supplement*, 30/3/09, p.14

Thomas, S., Peng, WJ. And Gray, J. (2007) Modelling patterns of improvement over time: value-added trends in English secondary school performance across ten cohorts, *Oxford Review of Education*, 33, 3, 261-295

Tymms, P. (2003) *School composition effects*, School of Education, Durham University, January 2003

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

Tymms, P., Merrell, C., Heron, T., Jones, P., Alborne, S. and Henderson, B. (2008) The importance of districts, *School Effectiveness and School Improvement*, 19, 3, 261-274