Commentary on the evaluation of teacher effectiveness through student test scores

Stephen Gorard

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

NESSE AHQ 36 - To what extent can Sanders' and Rivers' (1996 and subsequent) assertions about the impact upon student learning of teacher effectiveness be justified by the research they have undertaken and the methods they have used?

Summary

- The Tennessee Value-added Assessment System is claimed to be able to estimate the impact of teachers on their students' progress.
- This has led to further claims, such as that teacher quality is paramount in improving student progress.
- However, TVAAS and similar schemes should not be relied upon.
- Explanations of TVASS in the public domain are incomplete and poorly presented.
- TVAAS is not a 'test' of anything, and other analysts have attributed the same student progress residuals as used in TVAAS to school, classroom, district, leadership, social and economic factors.
- The analysis appears to be circular – effective teachers are defined by progress of students so students making progress have necessarily effective teachers.
- The analysis anyway cannot be sustained with the kinds of data available.
- The estimated level of missing data, and of measurement and representational error in the data that is present, suggest that the estimated residuals for each student are composed largely of error terms.
- Sanders and colleagues make elementary statistical errors, such as using significance tests with population data.

Introduction

Since at least 1996, Sanders and others (Sanders and Rivers 1996, Sanders and Horn 1998, Sanders 2000) have claimed to be able to estimate teacher effectiveness from student test scores. The claimed result is that 'Our research work… clearly indicates that differences in teacher effectiveness is [*sic*] the single largest factor affecting academic growth of populations of students' (Sanders, 2000, p.334). The Tennessee Value-added Assessment System (TVAAS) has been claimed to be 'an efficient and effective method for determining individual teacher's influence on the rate of academic growth for student populations' (Sanders and Rivers, p.1). It uses the academic test scores of students, tracked longitudinally, in a complex statistical analysis, to estimate the impact of teachers. There is a certain plausibility about their logic, which coupled with a hunger for teacher accountability measures, and a faith in technical solutions, has led some commentators to extol this approach. Barber and Mourshed (2007), for example, call the research by Sanders 'seminal' in showing how important effective teachers are, and how damaging poor teachers are, for student learning. They conclude that the quality of instruction in education is paramount, and therefore that the preparation of teachers is a key determinant of education quality. This research 'finding' is now reflected in some important policy documents, including those of the European Commission.

As with any logical argument the conclusion may be true even if the premises and intermediate steps are faulty. In what follows I say nothing about the likelihood of the conclusion. Yet every step in the argument towards that conclusion is questionable. Even if more or less effective teaching made a difference to student progress this would not necessarily make quality of instruction paramount. Even if quality of instruction were paramount this would not necessarily mean that teacher quality is determined by teacher education. And so on. But the key issue is whether the original premise about judging teacher effectiveness is valid. My answer to AHQ36 is that Sanders' (and Barber's) assertions about teacher effectiveness have almost no scientific justification. I start with a summary of the Sanders research, and the assumptions it makes.

Teacher 'effectiveness'

There are a number of reasons why policy-makers and education leaders might want to be able to evaluate the effectiveness of teachers, including for inspection, improvement, targeted development, incentive payments and, in extreme cases, dismissal. And for each of these reasons, policy-makers and education leaders might wish to specify a different version of teacher effectiveness. Teachers might be considered effective if they worked well together, could control their classrooms, or encouraged students to: attend school, select the teacher's bespoke courses, raise their occupational aspirations or stay in subsequent educational phases. Such teacher 'effects' might be immediate, as in inhibiting students from smoking at school, or longer-term, such as in inhibiting students from smoking in later life. Often, however, a very narrow and immediate definition of teacher effectiveness is used, focusing on what can be deduced about short-term learning from pencil-and-paper testing of students. The Tennessee Value-added Assessment System (TVAAS), for example, defines teacher effectiveness in terms of progress made by their students while at school, as judged by changes in their test scores.

For any set of schools or teachers, if we rank them by their student scores in assessments of learning, 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 of student attainment are not a fair test of school or teacher performance. TVAAS uses the 'scaled scores' of students (Sanders and Horn 1998, p.249) over time (usually an average of three years) in each curriculum area to calculate gain scores, also referred to as a student's progress.

Even at this level of generality, the model of teacher effectiveness makes a number of important assumptions. Among these are:

- The differences in student test scores in any stage of education can be attributed to the impact of teachers.

- The tests taken by students are a reasonably accurate measure of their teacher-directed learning.

- Teacher effectiveness is a relatively static phenomenon, and it is therefore appropriate to use past performance of students to judge the present, and perhaps future, effectiveness of teachers.

- Finding the difference between a prior and subsequent test score for each student yields a progress score which is then independent of raw-score levels of attainment.

Let us consider each assumption in turn

1. The differences in student test scores in any stage of education can be attributed to the impact of teachers.

Some early studies of school effectiveness famously found very little or no difference in the outcomes of schools once 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), and continue today (Lubienski and Lubienski 2006). The differences in student outcomes between teachers, 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 cases that can be explained (Shipman 1997, Tymms 2003). Looked at in this way, it seems that which teacher a student has, or which school they attend, makes little difference to their learning (as assessed by statutory tests).

However, over the past 30 or more years a different series of studies have come to an almost opposite conclusion, based on pretty much the same evidence. School and teacher effectiveness researchers accept that much or most of the variation in school outcomes is due to school intake characteristics (Rutter et al. 1979). 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). The TVAAS work follows this line of argument, except that it attributes the residuals to teacher effects alone. The idea that unexplained variation in student progress is attributable to teachers (or schools) is not tested by the modelling that ensues. It is taken on trust. What are the reasons this assumption might be false?

Perhaps most obviously, the residual variation in student gains scores has been attributed by analysts other than Sanders to other factors. These include external determinants such as the continuing influence of differential family support, socio-economic trajectories, and cultural and ethnic-related factors. They include school-level factors such as resources, curricula, timetabling and leadership. And they include educational factors beyond the school, such as district and areal policies and funding arrangements. Of course, all such attributions have no more justification that an attribution of the residual gain scores to the impact of teachers. But they are all in competition to explain the same small amount of variation (once prior attainment is accounted for). In addition, of course, the residual scores in VA calculations contain a substantial error component.

Sanders and Horn (1998) explain that they are dealing with 'fractured student records, which are always present in real-world student achievement data' (p.248). What they mean by this is that some student records will be missing or damaged, and some records that are present will contain missing data. They do not explain, in any of the sources I have been able to trace, how large a problem this is. In England, schools are annually required by law to provide figures for the National Student Database (NPD) on achievement and the Student-level Annual Schools Census (PLASC) on student detials. Both databases ostensibly have records for all students at school in England (but necessarily exclude any students not registered). NPD/PLASC is a high quality dataset, much better than any analyst would hope to generate through primary data collection, and yet missing data remains a substantial problem.

Independent fee-paying schools are not involved. So the PLASC/NPD dataset only includes 93% of the age cohort at best (minus also those educated at home, by other means, and some cases simply not registered at all). Around 10% of the individual student records are un-matched across the two databases. In 2007, the Key Stage 4 (15-year-old cohort) dataset contained records for 673,563 students. However, every variable, including the contextual and attainment variables, had a high proportion of missing cases. For example, at least 75,944 were missing a code for free school meal (FSM) eligibility (a measure of poverty). This represents over 11% of cases. Even when data is not coded as missing, it is effectively missing, such as the codes 'Refused' and 'Not obtained' which are additional to the missing data on student ethnicity. If we delete from the 2007 PLASC/NPD all cases that are unmatched, or missing FSM, in care, special needs, sex and/or ethnicity data, and at least one attainment score, then we end up with complete records for less than 60% of the school-age population.

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 student). So, analysts assume that that where we do not know when a student 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 score is an average attainer, and so on. These are very questionable assumptions. These kinds of assumptions have to be made in order not to lose all of those cases with at least one missing value in a critical variable. But making these unjustified assumptions then means that 40% or more of cases are very likely to have an incorrect value in at least one critical variable. There is no way that any kind of statistical analysis can make up for this (see below). And I simply do not believe that the school records for Tennessee in 1993 (broken down *also* in terms of the teacher for each subject – a clear area for the introduction of further errors) were more complete than this.

Of the information that is present in any schools database, some of it will be incorrect. 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). If we take the underlying competence of the student 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, over and above the errors caused by missing data.

The subsequent coding of 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 students every year. Then the data must be entered (transcribed) and low level errors are liable to creep in again. Data can even 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 student that is present and entered and stored 'correctly' is always liable to be in error, due to the change in number base and the finite number of binary digits used to store it in floating-point format in a computer or calculator.

Each of the two attainment scores in a teachers effectiveness model (and of course any other variables used such as the link between teachers and students) will have the kinds of errors illustrated so far. It would be conservative to imagine that a national or state assessment system was 90% accurate as a whole (or that 90% of students were recorded the correct mark/grade). It would also be quite conservative to imagine that, overall, only around 10% of the cases or variables used in a school effectiveness calculation were missing (or incorrectly replaced by defaults). This means that each

attainment score is liable to be no more than 80% accurate – or put another way the relative error is *at least* 20% in each set of figures used in an effectiveness calculation.

Such errors are said to 'propagate' through calculations, meaning that everything we do with our achieved measures we also do with their measurement errors. 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 $(X \pm \varepsilon_x) - (Y \pm \varepsilon_y)$. 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). I focus on subtraction here for two reasons. First, effectiveness models are at heart based on a gain score from one attainment period to another and this can be expressed as a subtraction. Second, in teacher effectiveness both of the attainment scores are positive (or zero). This means that X-Y (or whatever) will be smaller than X (and probably smaller than Y as well). So, finding the difference (gain) between X and Y reduces the number we use as our achieved measure (X-Y) while at the same time increasing the upper error bound by adding together the individual error components of X and Y. Put more starkly, the maximum relative error in the result increases – sometimes dramatically.

Imagine that one prior attainment score (perhaps a KS2 point score for one student in England) is 70 and that the subsequent attainment score for the same student (perhaps a KS3 point score) is 100. This gives a manifest gain of 30 points (from KS2 to KS3). Using the conservative estimates above, we could say that the first score, being only 80% accurate, actually represents a true figure somewhere between 56 and 84. The second true figure is somewhere between 80 and 120. Thus, under these assumptions our achieved estimate of the gain score, 30, really lies between -4 and 64. The relative error has changed from an estimated 20% in each of the original figures to well over 100% in our computed answer. Subtracting two positive numbers of a similar order of magnitude dramatically increases the relative error bounds in the answer, and this applies whatever the relative error was in the original figures. And whatever the size of the initial error it tends towards infinity as the gain score for any student decreases towards zero – or put another way the smaller the gain score the less accurate it is (because the relative error in a score of zero is infinite for any finite initial absolute error).

In summary a very high proportion of the apparent gain scores for any student will actually be an error component deriving from the propagation of missing data, measurement errors, and representational errors. It would be quite unwise to attribute the meaningless differences in these 'scores' to the influence of teachers.

2. The tests taken by students are a reasonably accurate measure of their teacher-directed learning.

Most of the variation in gain scores between students will be the result of error. There may be a small 'residual' of this residual that could be attributed to the impact of teachers (and of course to all other competing explanations such as the continuing effect of student background). But it is hard to see how this might be identified separately and quantified in practice.

Not all areas of teaching are routinely subject to statutory testing in Tennessee or elsewhere (Sander and Horn 1998). Even in England which has a famously prescriptive programme of statutory testing at ages 7, 11, and 14, the focus is largely on maths, science and English. This means that some teachers cannot be included anyway since their subject contributions are not tested for (most obviously perhaps sports and PE staff).

It is very rare for one student to come into contact with only one teacher, even for one subject. Team-teaching, teaching assistants, on-line and virtual participation, and replacement and student teachers, among other factors, will confuse the issue. Teachers and their styles might vary over time, and might be effective for some students but not others. Their effectiveness might depend on the precise topic taught.

3. Teacher effectiveness is a relatively static phenomenon, and it is therefore appropriate to use past performance of students to judge the present, and perhaps future, effectiveness of teachers.

Our lack of ability to calibrate the results of school effectiveness models against anything except themselves is a problem. In everyday measurements of time, length, temperature and so on we get a sense of the accuracy of our measuring scales by comparing the measurements with the qualities being measured (Gorard 2009). There is no equivalent for teacher effectiveness (if we had a true or direct measure of teacher effectiveness then we would not need Sanders' technique anyway). The scores are just like magic figures emerging from a long-winded and quasi-rational calculation. Their advocates claim that these figures represent fair 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 calculations did not work. What would we expect to emerge from teacher effectiveness studies? 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 teachers/schools to be the same, once 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. This is a common symptom of pseudo-science. 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 effect (Tymms 2003).

If the VA residuals were mostly due to error, how would the results behave? We would expect the 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 teachers and schools will have apparently consistent positive or negative VA 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 VA is zero-sum by design, around half of all schools and teachers in any one year would have positive scores and half negative. If the VA were truly meaningless, then we might expect around one quarter of all schools to have successive positive VA 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, et al. (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 over 3 years means that only 25% of the variation in VA is common to all years. Is this 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, or does it just illustrate again that VA is very sensitive to the propagation of relative error?

These authors were largely dealing with 'school' effectiveness, analysing results for an entire cohort and treating all subject areas as equivalent for analytic purposes. This means that the average number of cases per school might be 100 or more. In teacher effectiveness on the other hand, which attempts to measure progress in terms of individual school subjects and teachers, the largest number of cases involved is likely to be a teaching group of around 30 students or less. Irrespective of all other factors, this will make teacher effectiveness scores much more volatile even than purported school effects, because of the small numbers involved.

Of course, the process tells us only what the teacher had been like and not what they will be like. But the coefficients in VA models, 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 2008a). With enough variables, combinations of variables and categories within variables it is possible to create a perfect (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 measures were completely different – probably because 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).

4. Finding the difference between a prior and subsequent test score for each student yields a progress score which is then independent of raw-score levels of attainment.

The key calculation underlying school and teacher effectiveness is the creation of the residual between actual and predicted student scores (or between prior and posterior scores). Since this is based on two raw scores (the prior and current attainment of each student), it should not be surprising to discover that value-added (VA) results are highly correlated with both of these raw scores (Gorard 2006, 2008b). Around 50% of the variance in gain scores is common to the prior score (Pearson R of over 0.7), and 50% to the posterior score. In fact, the correlation between prior and current attainment is of the same order as the correlation between prior attainment and VA scores. Put more simply, VA calculations are flawed from the outset by not being independent of the raw scores from which they are generated. They are no more a fair test of performance than raw scores are.

The irrelevance of technical solutions

It is worth pointing out at this stage that any analysis using real data with some combination of (almost) inevitable measurement errors will be biased, and so 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 of these sources of error lead to random measurement error (of the kind that might come from random sampling variation, for example). Those without test scores, 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. Similarly, representational errors in denary/binary conversion are part of the numbering systems involved and entirely predictable (given enough time and care). Like every stage in the error generation process described so far, they are not random in nature, occurrence or source.

Therefore, the relative error bounds illustrated above do not represent likelihoods, or have any kind of normal distribution, because the errors are not random in nature. An error of 100% is as likely as one of 50% or zero. And for any one school, region, key stage, subject of assessment, examination board, or socio-economic group all (or most) of the errors could be in the same direction. There is no kind of statistical treatment based on probability theory that can help overcome these limitations.

Whether as simple as confidence intervals or as complex as multi-level modelling, such techniques are all irrelevant.

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. But when working with population figures such as NPD/PLASC these techniques mean nothing. 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. Sanders and Rivers (1996) state quite clearly that they are working with the 'entire grade 2-8 student population' for Tennessee (p.1). Yet their reported analysis cites statistical significance, p-values and F-statistics calculated for this population (e.g. p.3). These are statisticians who do not understand basic statistical principles. This kind of work has been described as 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.

It is also important to recall that VA is a zero-sum calculation. The VA for a student, 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 teacher 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.

Conclusion

In addition to the concerns raised in the memo attached to AHQ36 – the work is not causal, attempts no intervention, and is retrospective rather than longitudinal in the true sense – this consideration has illustrated the propagation of initial error. There is no way to avoid it. A measurement with a 20% margin for error is frequently usable in social science, but an answer with a 100% margin is generally useless. It is certainly no basis for making policy, rewarding heads, informing parents, condemning teachers, or closing schools.

Sanders and Horn (1998, p.254) claim that 'African American students and white students with the same level of prior achievement make comparable academic progress when they are assigned to teachers of comparable effectiveness'. What does this mean? The teacher effectiveness is calculated on the basis of the progress made by students, so this claim by Sanders and Horn is completely tautological. In fact their whole argument about the importance and impact of teachers is circular. Effective teachers are defined as those with students making good progress, so obviously, but

by definition only, students make good progress with effective teachers. Empirically, this means nothing, even when dressed up in the most complex of (invalid and exclusionary) statistical analyses. One has to consider whether incompetence or something worse lies behind the ill-thought claims of Sanders and others.[1]

The teacher effectiveness model does not and could not work as intended. No reliance should, in my opinion, be placed in it. Perhaps more importantly, once policy-makers understand that they cannot legitimately use this approach to differentiate teacher performance, they may begin to question the dominance of the school effectiveness model more generally.

References

Barber, M. and Mourscheд, M. (2007) *How the world's best-performing school systems come out on top*, McKinsey & Co.

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

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

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

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

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

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

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

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

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

[1] Of course, it is possible that Sanders and colleagues have explanations for these seemingly unanswerable issues. If so they are keeping them very quiet. In a sequence of papers (in the same journal) Sanders has cross-referred and made substantial claims for the system but never once explained in detail how it works and why it is not circular. I have followed the chain of references, citing each other as giving more details, as far as I am able and have found nothing. What is there is poorly and incompletely explained.

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

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

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

Sanders, W. and Rivers, J. (1996) *Cumulative and residual effects of teachers on future student academic achievement*, University of Tennessee: Value-added Research and Assessment Center

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

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