What difference do teachers make? A consideration of the wider outcomes of schooling

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

Abstract

This paper is based on a series of previous research studies looking at the impact of teachers in the UK and internationally. It suggests that there is no way to demonstrate with convincing evidence that some teachers are more or less effective with equivalent pupils, in terms of test outcomes. This is not necessarily because teachers are not differentially effective, but because the calculations involved are not possible with our current methods. And, of course, this is not to suggest that teachers in general do not make a difference – only that they are not obviously *differentially* effective. However, there is indicative evidence, particularly from pupil reports, that there is considerable variation in the skills, even very basic skills, of teachers. This may be partly the result of variability in the process of admitting and qualifying trainee teachers. There is also good evidence that the quality of pupil:teacher interaction in schools is linked to pupils' sense of justice, trust in others, and reports of citizenship activity. Perhaps then this is the clearest difference that individual teachers make, through their daily behaviour and on the wider outcomes of schooling.

Introduction

Attracting, retaining and developing teachers of the highest quality are concerns for the education systems of all countries. A common belief among policy-makers and many other commentators is that there is considerable variation in the effectiveness of teachers, and that this differential effectiveness can be measured and so rewarded or penalised. Such effectiveness is almost always conceived in relation to pupil academic attainment at school. Good teachers are, in this view, those that teach the pupils who then gain the best possible test results. Of course, when pressed, such commentators would concede that this approach to identifying differentially effective teachers is not easy. Confounding factors include the background and prior experiences of the pupils, the initial talent of the pupils, the variability between alternative measures of attainment in terms of examining body, year, syllabus, region, mode of examination, and subject, and the inconvenient fact that most pupils are taught by more than one teacher, perhaps including those outside the school system such as family, peers, and tutors. In order to overcome all of this 'noise' the differential impact of teachers on their pupils' attainment would have to be very considerable in order to be safely identified. Yet there are systems in operation around the world that claim to be able to make that identification safely, and teachers are therefore rewarded and punished on that basis. This paper considers such systems, focussing in particular on the long-running Tennessee Value-added Assessment System (TVAAS) in the USA, and suggests that it is not currently possible to identify a differential teacher effect on pupil attainment or progress. However, there is good evidence that not all teachers

have the same levels of pedagogic competence, in the view of their pupils. The paper then considers some of the ways in which we might currently be able to demonstrate that teachers do have a differential impact on their pupils. The paper concludes by considering the implications of this argument, if accepted.

Sources of evidence

This paper is based on four separate major sources of evidence. There is insufficient space here to rehearse each of these in detail, so a brief summary is provided for each, with references to the full methods used, that appear in the public domain as peer-reviewed publications.

The evidence relevant to the critique of differential teacher effectiveness is largely drawn from a study conducted for the EU Directorate General for Education and Culture, and is based on an analysis of the National Pupil Database (NPD) in England (Gorard 2010a). The NPD is an excellent resource for studies of school and teacher effectiveness containing longitudinal records for all pupils in England including their background, socio-economic characteristics, prior attainment, courses taken and qualifications obtained. The discussion of the quality of teacher preparation comes from a number of studies of teacher supply, funded by the General Teaching Council of Wales, the Teacher Training Agency (as it then was) and the Economic and Social Research Council, as summarised in Gorard et al. (2006). This involved, among other things, a re-analysis of the intakes and outcomes for all teaching training courses in England over a number of years. The evidence on pupils' views of their teachers' competence comes from a large-scale national study for the (then) Qualifications and Curriculum Authority intended to establish a baseline picture for the 14-19 curriculum reforms in England (Gorard and See 2011). The study involved a survey of 4,900 pupils, and 798 pupils in focus groups, from years 11 and 12. The evidence on wider outcomes of schooling and the importance of pupil:teacher interactions comes from a series of studies in six EU countries plus Japan, funded by the EU Socrates Programme. This included a survey of nearly 20,000 pupils across all countries (Gorard and Smith 2010). Combining the results of these studies here for the first time provides an interesting, and somewhat controversial, conclusion.

Teachers and attainment

It is generally accepted that pupils are not an entirely blank-slate, certainly not by the time of their second and subsequent teachers. Pupils have some kind of pre-existing propensity to do better or worse in standard tests of attainment at school. This could be called talent and/or motivation, but why it exists, if it does, and whether it matters is not the subject of this paper. The significance for this paper is that it is widely recognised that it would not be fair to judge the proficiency of teachers solely on the final results of their pupils, because some pupils are intrinsically easier to teach than others. To take an obvious example, it would not be fair to conclude in a selective system of schooling that teachers in the grammar schools are more competent than those in secondary-modern schools simply because pupils in grammar schools tend to perform better in measures of attainment. The pupils have been selected at age 11 to attend grammar schools precisely because they appear to be the most likely to do well

in similar tests at age 16, for example. The fact that this selection is successful, in its own terms, does not provide any evidence at all about relative teacher competence. What is needed instead, it appears, is a method of separating the progress made by pupils that is attributable to their teacher from their pupils' overall level of attainment, from the progress attributable to other influences (including, of course, other teachers), and from any errors in the measurements. This 'value-added' approach sounds challenging, and indeed it is so.

To see how problematic this more sophisticated approach is in practice, the paper uses the Tennessee Value-added Assessment System (TVAAS) as an important case study. 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 pupil 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 pupils' (Sanders, 2000, p.334). TVAAS has been claimed to be 'an efficient and effective method for determining individual teacher's influence on the rate of academic growth for pupil populations' (Sanders and Rivers, p.1). The system uses the academic test scores of pupils, tracked longitudinally, in a complex statistical analysis, to estimate the impact of teachers. There is a 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 pupil 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 McKinsey report, and others following, has been highly influential, and the research 'finding' about the importance of 'good teachers' is now reflected in some important policy documents, including those of the European Commission and the OECD (Coffield 2012).

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 pupils 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 pupils from smoking at school, or longer-term, such as in inhibiting pupils 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 pupils. TVAAS, for example, defines teacher effectiveness in terms of progress made by their pupils while at school, as judged by changes in their test scores.

For any set of schools or teachers, if we rank them by their pupil scores in assessments of learning, then we would tend to find that schools at the high and low ends differed in more than their pupil assessments. Schools in areas with more expensive housing (or more local income in the US), schools that select their pupil 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 pupil 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 pupil attainment are not a fair test of school or teacher performance. TVAAS uses the 'scaled scores' of pupils (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 pupil's progress.

Even at this level of generality, the model of teacher effectiveness makes a number of important assumptions. The paper considers each of these assumptions in turn

Differences in test scores can be attributed to the impact of teachers?

Some early studies of school effectiveness found little difference in the outcomes of schools once their pupil intake differences had been taken into account (Coleman et al. 1966). Such studies, using either or both of pupil prior attainment and pupil family background variables, are still being published (Lubienski and Lubienski 2006). The differences in pupil outcomes between teachers or individual schools can be largely explained by the differences in their pupil intakes. School and teacher effectiveness researchers accept that 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 pupil intake) is evidence of differential teacher or 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 pupil progress is attributable to teachers (or schools) is not tested by the modelling that ensues. It is merely taken on trust. What are the reasons this assumption might be false?

Perhaps most obviously, the residual variation in pupil gain scores has been attributed by analysts other than Sanders to other factors. These factors 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 than 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 contain a substantial error component.

Sanders and Horn (1998) explain that they are dealing with 'fractured pupil records, which are always present in real-world pupil achievement data' (p.248). What they mean by this is that some pupil records will be missing or damaged, and some records that are present will contain missing data. They do not explain, in any publication how large a problem this is.

By comparison, in England, schools are annually required by law to provide figures on pupil characteristics and qualifications for the National Pupil Database (NPD). The

database ostensibly has records for all pupils registered as being at school in England. The NPD 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 even here. Around 10% of the individual pupil records are un-matched across years and phases of schooling. In 2007, the Key Stage 4 (15-year-old cohort) dataset contained records for 673,563 pupils. However, most variables had a high proportion of missing cases. For example, at least 75,944 were missing a code for free school meal eligibility (a measure of poverty). This represents over 11% of cases. Once all cases that are unmatched, or missing a key background variable or an attainment score, are deleted from the 2007 NPD, then there are 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. So, analysts assume 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, and so on. These assumptions are made in order to retain most cases with at least one missing value in a key variable. But statistical analysis cannot make up for this, because the missing values are heavily biased rather than random (Gorard 2012a, and see below).

What is true for NPD in 2007 and subsequently will also be true for the school records for Tennessee in 1993 and subsequently. Note that in order to judge teacher effects, Sanders had to break down these pupil figures *also* in terms of the teacher for each subject. This is a clear area for the introduction of further errors.

Some of the information that is present in any schools database, whether NDP or TVAAS, will be incorrect. Assessment via examination, project, coursework or teacher's grading is an imperfect process (Nuttall 1979, Newton 1997). Learning is not an easy thing to measure, unlike height or the number of pupils in a classroom. However well-constructed the assessment system, there will be considerable measurement error, over and above the errors caused by missing data. Then there will be a small number of additional errors stemming from coding, transcription, matching and storage of records.

Each of the two attainment scores in any teacher effectiveness model (and of course any other variables used, such as the link between teachers and pupils) 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 pupils were recorded with 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 done with these scores is also done with their measurement errors. Since we do not know whether the error in either score is positive or negative when we subtract are effectively adding the error components. This means that the estimated gain score for each pupil is the difference between the predicted and achieved score (obviously a number that is considerably smaller than either score) plus the sum of the errors in both scores. Put more simply, the error will tend to get larger and the number in which

the error occurs will get a lot smaller. A realistic 20% relative error in both the predicted and achieved scores for each pupil could lead to a relative error of 10,000%, and often much more, in the pupil gain score. This makes the gain scores completely unusable in practice.

In summary, a very high proportion of the apparent gain scores for any pupil 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.

Teacher effectiveness is a relatively static phenomenon?

The fact that teacher effectiveness data contains substantial 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 researchers claim as evidence of effectiveness, the more this could also be evidence of the error component. 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, Kelly and Monczunski 2007). There is huge year-on-year variation in purported school effects (Gorard 2011a). School effectiveness work involves 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 pupils or less. Irrespective of all other factors, this will make teacher effectiveness scores even more volatile than purported school effects, because of the small numbers involved. 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 so two VA measures were calculated. The measures were completely different – probably because they did not really mean anything at all!

If teacher effectiveness, as far as we can judge it, is a volatile phenomenon altering easily across years, classes and pupils, then it would not be appropriate to use the past performance of pupils 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 therefore 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 2010b). There is no equivalent for teacher effectiveness. Advocates claim that effectiveness figures represent fair performance measures, but they can provide nothing except the purported plausibility of the calculation to justify that.

Tests taken by pupils are an accurate measure of teacher-directed learning?

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 in any teacher effectiveness system, since their subject contributions are not tested for (most obviously perhaps sports and PE staff).

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

Therefore, much of the variation in gain scores between pupils 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 pupil background). But it is hard to see how this might be identified separately, and then quantified in practice.

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 pupils 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. There is no kind of statistical treatment based on probability theory that can help overcome these and other limitations. Whether as simple as confidence intervals or as complex as multi-level modelling, such techniques are all irrelevant.

Unfortunately the field of teacher effectiveness research works on the invalid assumption that errors in the data are random in nature and so can be estimated and corrected, based on techniques based on random sampling theory. But when working with population figures such as NPD 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. Sanders and Rivers (1996) state quite clearly that they are working with the 'entire grade 2-8 pupil 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.

The progress score for each pupil is independent of level of attainment?

Sanders and Horn (1998, p.254) claim that 'African American pupils and white pupils 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 pupils, so this claim by Sanders and Horn is tautological. In fact, their whole argument about the importance and impact of teachers is circular. Effective teachers are defined as those with pupils making good progress, so obviously, but by definition only, pupils make good progress with effective teachers. Empirically, this means nothing.

The key calculation underlying school and teacher effectiveness 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 value-added (VA) results are highly correlated with both of these raw scores (Gorard 2006, 2008). 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 actually no more a fair test of performance than raw scores are.

Evidence of poor teaching

If it is therefore not yet possible to identify differentially effective teachers because of the confounding factors and errors in the measures of pupil progress, what does that mean for the strong intuition that some teachers are better than others? After all, absence of evidence is not evidence of absence. There may be a teacher effect on pupil learning but it may be too small in comparison to the noise in the measuring system. It is clear that pupils themselves report considerable variation in the quality of teaching, as they experience it (Gorard and See 2011).

Pupils appreciate a good relationship with their teachers, which for young people often means being in a relationship of mutual trust and respect. For some pupils it is very simple things like addressing their teachers by their first names that make them feel they were being treated as grown up:

> Teachers are much, they respect you more, talk to you like, not like you're a little kid, treat you with a bit of respect, give you a bit of leeway if you're like that with them, if you do what they do, they'll be alright with you. They won't talk to like a little child or look down at you or anything, so that's cushedy.

This is in stark contrast to situations where pupil-teacher relationships can be a cause of stress for pupils:

> Some teachers in the school respect certain pupils and don't respect some others, and they wonder why kids get so rude to them and start swearing and that's when we get into trouble. The teachers say we want respect from you but they don't normally show it to us. They're the teacher they're always right, we're the kid and we don't know what we are going on about.

There are even accounts of teachers mocking their charges in a manner likely to depress aspiration:

> When I actually went for an interview at [agricultural college] and got my place... I showed my [pastoral teacher] and… she actually turned round and said that it's a load of rubbish, there's no point doing it, 'cos I ain't going to get nowhere in life 'cos I never come to school… I might as well just drop all my dreams and just be a bum, basically, live off Social. Which really put me down.

Pupils appreciate innovation and preparation in learning activities. Some pupils said they enjoyed lessons which involved physical activities like getting them to move around or acting out a scene. Others liked practical work, debates, dramatisations or just the unexpected – almost anything where there is variation in delivery and activity:

> We had to do this electrons thing and then the teacher got us all to stand round in a circle and we had to hold this rope, and we were the electrons and then we kind of moved around, so they only move in one way the electrons. If like one falls, then the other can like push the other one.

Again, there are pupil accounts of teacher behaviour in stark contrast. In fact, poor basic pedagogic skill and uninspiring teaching were the basis for most complaints from pupils. There were accounts of having to listen to a dull teacher for lengthy periods, passive copying, note taking, and having to sit still:

> He just stands at the blackboard, or the whiteboard…and just writing on the board…

> [The teacher will] give you a sheet and you'll just go through each question and they might not even, unless you express a want to be told how to do it, they might not even, you know, support you in any way.

A common complaint was about audibility – surely one of the most basic elements of classroom craft:

> In (subject) teacher can't speak.

> I am failing my (subject) because my teacher does not speak up and only talks to the front row.

A linked complaint was lack of real engagement, such as eye contact, and continuing with a narrative almost like a recording, or turning question and answer sessions into the cliché of comedy shows:

> Teachers often go too fast, like you're saying; you don't end up knowing what's going on because you're trying to take notes like as fast as you can, and so you are not listening, so then your notes aren't completely there, so …

> He doesn't say can you answer this question most of the time, because he just shouts.

> Cos, as soon as the teachers ask a question he doesn't like give anyone else a chance to think about it, he just shoot off. Basically all of you just sit there, you're still thinking about the question and (expletive) gives the answers.

How is this possible?

According to pupils there is clearly huge variation in the quality of their teachers (in England) and yet we are seemingly unable to measure the difference this makes to pupil attainment. If there is such variation, how could it arise? One possibility suggested by previous evidence is that some of the better candidates for teacher training in England are being turned away while some others are accepted, because of the institutional nature of the application process (Gorard et al. 2006). Another, related, possibility is that teacher trainees are being passed in some institutions who would fail in others. This is because a quota of places is awarded to each teacher training institution, and they have the responsibility for accepting trainees. More trainees apply every year than are accepted, but those rejected by some high prestige institutions are actually better qualified than the best qualified trainees accepted at some lower prestige institutions. Only a more co-ordinated national system of application and selection can prevent this. The problem of such variability in teacher quality may also arise because each training institution is permitted to decide on trainee outcomes. There is no relationship at all between the intake qualifications of trainees and their outcomes. This might be because the institutions with less qualified intakes are genuinely performing better, and so levelling the playing-field for their trainees. However, OFSTED inspections also suggest that several of these institutions just do not have the capacity to judge teacher quality for themselves, even though they are responsible for qualifying teachers. What damage could this be doing?

The wider outcomes of schooling

Given the variation in pupils' reported experiences of teaching and of their relationships with teachers, it would be odd to find no evidence of any impact of this on the pupils. What follows is based on a set of associations rather than a definitive test of influence, but in that it is no different to studies of teacher academic effectiveness, which are also purely correlational. What follows differs from traditional studies of school and teacher effectiveness because the 'effect' sizes uncovered are far larger (large enough to overcome noise in the measurements), perhaps because they are much less stratified by pupil background characteristics. Therefore, it is reasonable to conclude that what follows is indicative evidence of the potential impact of teachers on wider school outcomes, such as whether pupils learn to trust their teachers, and so trust others in wider society, and whether they are willing to forego something in order to help others (Gorard 2011b, 2012b).

Pupils' reported experiences of interacting with their teachers are generally consistent across social, economic and family background groups. In all countries covered, males, females, high and low attainers, those from families with professional educated parents and those with less educated or unemployed parents, recent immigrants and

second language speakers, for example, all report pretty much the same experiences. This lack of stratification is rare in the sociology of education.

Around 37% of the variation between pupils reporting they consider their teachers to be trustworthy, or not, is accounted for by their experience of interactions with teachers. This is not surprising. Only around 14% of the variation is accounted for by pupil background (compare this with the 80% to 90% of variation in test scores explained by pupil background alone in teacher effectiveness studies). Pupils who report receiving school marks that fairly reflect the effort they made, and the quality of their work, are much more likely to trust their teachers. More importantly, teachers are more likely to be trusted when they are seen to punish poor behaviour appropriately. So teachers must be prepared to reward and punish some pupils when this is warranted, and to remember not to carry this discrimination over into areas of school life, where it is not warranted. Even more importantly, teachers have to respect all pupils, show concern for all, and encourage individual pupil autonomy, in order to be trusted. These pupil accounts are very similar to those above about the mixed quality of teaching, and how enjoyable or otherwise lessons are.

A similar, if slightly weaker, picture appears when pupils report whether they consider people more generally, and outside school, to be trustworthy. Around 9% of the variation is accounted for by pupil background, whereas 13% is accounted for by school experiences, including pupil interactions with teachers. To a large extent, the same factors are involved. Pupils tend to trust people more generally when their teachers are reported to have been trustworthy, by using marks and punishments fairly and treating all pupils the same otherwise. This suggests a clear role for teachers in educating citizens who are generally trusting of others. They do this through their exemplification of good (or indeed poor) citizenship in action. And the result is associated not only with better pupil relationships but also more hopeful pupil engagement outside and beyond schooling.

Again a similar pattern appears when pupils report, via a vignette episode, whether they would be willing for someone else with a specific difficulty to be helped at the cost of teacher attention to the pupil themselves. Around 18% of the variation is accounted for by pupil background and school factors, whereas 17% is accounted for by reported interaction with teachers. The latter includes whether teachers were interested in pupil well-being or whether they got angry with a pupil in front of others. This body of evidence, and more like it, suggests a possible role for teachers in educating citizens who are participatory, tolerant and supportive of the difficulties of others. Teachers appear to do this not only through citizenship pedagogy but through their exemplification of good citizenship in action.

Conclusions

When assessing the impact of teachers on pupil attainment, the propagation of initial error and the stratified nature of the confounding variables faced are such that no teacher ‘effect’ can be safely attributed. The teacher effectiveness model does not work as intended, and should not be used for making policy, rewarding heads, informing parents, condemning teachers, or closing schools (Gorard 2010a). This is in no way an argument for using raw-scores to assess teacher performance instead. If

anything, that would be even worse. Nor does it suggest that teachers, in general, are not useful and important. It is simply that with our current approaches based on pupil test outcomes we cannot safely identify any individual teacher who is differentially effective with equivalent pupils.

Despite this, the belief persists that some teachers are very different to others, and the reports of their pupils confirms this. There are accounts of excellent, imaginative, and co-operative learning. And there are, at the other extreme, reports of individual teachers whose lack of basic pedagogic and social interaction skill is so great that it is hard to envisage how they function as teachers at all. Their existence, if these pupil voices are accepted as valid, suggests that a lot of work needs to be done in attracting, selecting and developing appropriate teachers (Gorard et al. 2006). Possible improvements for England and similar countries would be to have a national system of selection and admission into teacher training, and to set up in-career development schemes and inspections to remind practicing teachers of the lasting importance of their behaviour when interact with pupils.

All young people must interact with teachers, and for an extended period. Wider school outcomes, derived more from schools as societies than from curriculum content, are therefore more likely contenders to provide evidence of teacher impacts at present. This is because although the errors in measuring these 'soft' concepts will be even greater than in measuring pupil performance, these errors do not propagate in the same way as they do in a value-added calculation. Also the scale of the differences between pupils is larger than for academic test scores, once pupil background is accounted for, because the outcomes are much less stratified by prior attainment, sex, parental origin and so on.

Overcoming the differences in school experiences reported by pupils does not require new buildings, extra teachers, or different kinds of schools. The differences should be relatively easy to address, in comparison to stubborn inequalities in attainment based on heavily stratified underlying variables. Pupils learn to trust others partly as a consequence of how trustworthy others have appeared to be so far in their lives, for example. Those for whom school was fair, and their teachers were just, were much more likely than others to report trusting the government of their country and most people in general. Pupil experiences at school and in interaction with their teachers can be envisaged as part of the creation of pupils' views on what a fair world would be like and whether a it is possible. If so, teachers, leaders and policy-makers have a direct responsibility to assist pupils in making positive but appropriately critical judgements. Note that this is not primarily a pedagogical or curriculum issue. Pupils learn about what society is like through their lives at school. There is little point in overtly teaching that people can be trusted if pupils are not trusted in schools, and if teachers do not behave according to what pupils see as the basic principles of justice.

References

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

Coffield, F. (2012) Why the McKinsey reports will not improve school systems, *Journal of Education Policy*, 27, 1, 131-149

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

Gorard, S. (2010a) Serious doubts about school effectiveness, *British Educational Research Journal*, 36, 5, 735-766

Gorard, S. (2010b) Measuring is more than assigning numbers, pp.389-408 in Walford, G., Tucker, E. and Viswanathan, M. (Eds.) *Sage Handbook of Measurement*, Los Angeles: Sage

Gorard, S. (2011a) Now you see it, now you don't: School effectiveness as conjuring?, *Research in Education*, 86, pp.39-45

Gorard, S. (2011b) The potential determinants of young peoples' sense of justice: an international study, *British Journal of Sociology of Education*, 32, 1, 35-52

Gorard, S. (2012a) Who is eligible for free school meals?: Characterising FSM as a measure of disadvantage in England, *British Educational Research Journal*, http://dx.doi.org/10.1080/01411926.2011.608118

Gorard, S. (2012b) Experiencing fairness at school: an international study in five countries, *International Journal of Educational Research*, 3, 3, http://dx.doi.org/10.1016/j.ijer.2012.03.003

Gorard, S. and See, BH (2011) How can we enhance enjoyment of secondary school?: the student view, *British Educational Research Journal*, 37, 4, 671-690

Gorard, S. and Smith, E. (2010) *Equity in Education: an international comparison of pupil perspectives*, London: Palgrave

Gorard, S., See, BH., Smith, E. and White, P. (2006) *Teacher supply: the key issues*, London: Continuum

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

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 pupil 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 pupil academic achievement*, University of Tennessee: Value-added Research and Assessment Center