Testing Teachers:

What works best for teacher evaluation and appraisal

March 2013

Richard Murphy

Improving social mobility through education

THE SUTTON TRUST

Contents

Foreword

Good schools are essential if we are to make the most of the talents and abilities of all our young people. There is now widespread acceptance among researchers within the UK and internationally that good teaching is at the heart of good schools, and must therefore be at the heart of any school improvement programme.

In recent years, both the Coalition and Labour governments have focused heavily on improving the quality of new teachers entering the profession. Programmes like Teach First, the Graduate Teacher Programme and more recently, School Direct, and a series of Golden Hello and bursary schemes have improved the standing of teaching as a profession and encouraged more good graduates to consider teaching as a career.

But, with 440,000 teachers in English classrooms, and 35,000 new teachers recruited each year, it is not enough simply to raise the quality of new teachers. It is more important to raise the standard of those already in the classroom, many of whom will be working with young people for decades to come.

There have already been significant changes in the flexibilities open to academies and other schools in how they appraise and evaluate teachers. Appraisal has been freed up. All schools are likely to have the chance to link pay for teachers more closely to their performance in the classroom rather than length of service in the future.

When the Labour government first introduced performance related pay in the late 90s, it did so within a very bureaucratic framework that failed to achieve its goals of linking extra rewards to the best performance in most schools. Michael Gove is removing many of those restrictions, and is hoping that doing so will mean schools feel free to use appraisal and evaluation to achieve real improvement and reward the best teachers more effectively. A Sutton Trust survey of teachers last year showed growing support for doing so.

But unless schools and their leaders develop their own clear appraisal standards, there is every danger that the extra freedoms will be no more effective than what went before. There is now much more powerful research on effective evaluation than ever before, and that's why this report from Richard Murphy from the London School of Economics for the Sutton Trust is so important. He has looked at the latest evidence from the US and UK on teacher evaluation and produced a useful analysis and guide that should help schools, and their leaders and governors, to devise systems that are fair and effective in a rapidly changing educational environment.

Earlier research for the Sutton Trust has shown that if we were to raise the performance of the poorest performing tenth of teachers to the average, we would move into the top rank of the PISA tables internationally. But there is a more compelling reason to do so: by improving the quality of our teachers collectively, we can ensure that every child has a decent education, and is not held back by poor teaching. That is a goal collectively worth pursuing.

I am very grateful to Richard Murphy for his work on this report. It will be one of a number of research inputs to be discussed at a summit on teaching, which we are jointly organising with the US based Foundation for Excellence in Education.

Sir Peter Lampl
Chairman
The Sutton Trust and the Education Endowment Foundation

Executive Summary

The increasing flexibility enjoyed by academies and other schools over teachers' pay and changes in the appraisal regulations in 2012 give schools in England a real opportunity to shape teacher evaluation and development to improve standards and reduce in-school variations between subjects and between pupils of different backgrounds.

The OECD (2009) concluded that *"the effective monitoring and evaluation of teaching is central to the continuous improvement of the effectiveness of teaching in a school".* Yet how this is achieved has still to be resolved. There is growing evidence from the United States and this country showing that there is a significant correlation between teacher evaluations and exam results. However, the evidence also suggests that schools should rely on a combination of approaches to gain a fuller picture of teacher effectiveness, and that teachers should be assessed on their cumulative performance over several years rather than on the data from a single year.

What is also clear is that effective evaluation is good for pupils and good for teachers. It can improve the quality of teaching, provided it is accompanied by good feedback, and it can lead to better results for pupils and improved learning.

It is important that schools use a clear approach to appraisal that is well understood by every teacher, and that they provide effective training for any staff members involved in evaluation. Using distinct appraisal and developmental systems with common standards will encourage honest feedback which is key to development. There can be value in using external expertise both to develop an effective approach and to benchmark standards.

England's decentralised evaluation system allows for discretion when making decisions based on these measures. A centralised decision-making process that is prescriptive will undoubtedly lead to cases of misclassification, given the impreciseness of these measures. Teacher evaluation metrics are not absolute and therefore they should only be used as indicators of performance. We must rely on the expertise of experienced school leaders to make informed decisions when appraising a teacher, taking all factors into account including those that impact on achievement and the strengths of each measure.

This decentralisation also means teachers' activities outside the classroom can be considered. Schools are complex working environments and a teacher's contribution to effective management and extra-curricular activities is also important.

Ways to evaluate teacher effectiveness

The three most common ways to evaluate teacher effectiveness are gains in test scores, classroom observations and pupil surveys. Each method has weaknesses, but each has its place within a comprehensive teacher evaluation system.

Gains in test scores for teacher performance: Gains in pupil test scores are the best available metric to measure teacher performance. Improvements in student attainment may be an imperfect measure, but they are a good starting point. The main advantage of this measure is its objectivity; despite its shortcomings, it is by far the most reliable of the three measures in predicting a teacher's future performance. Test and exam results cannot reliably be used to differentiate teachers who are just above and below average, but they can effectively be used to identify teachers who consistently perform well or badly. Schools in England are ideally placed to implement this as national tests and the Key Stage achievement levels provide common measures of attainment across subjects, schools and time.

Classroom observation for teacher development: Even when conducted by well-trained independent evaluators, classroom observations are the least predictive method of assessing teacher effectiveness. However, being

observed does allow for an unrivalled opportunity to provide constructive feedback to teachers. To promote honesty in the feedback, developmental and evaluative observations should be carried out separately. Observations are common in schools in England today but, for them to be most effective, clear standards must be established. Again, schools in England have standardised measures of teacher performance that can be used to this effect.

Pupil surveys for corroborating measures: Whilst pupil surveys are open to accusations of misreporting by pupils, it has been found that they do contain information on the effectiveness of the teacher. Student surveys are not as predictive as test score gains, and nor do they provide as much effective feedback as peer observation, they do provide a middle ground, against which, gains in test scores and classroom observations can be calibrated.

No measurement is perfect; all measurements are vulnerable to irrelevant factors and could be driven by outliers. However, with knowledge of their shortcomings, we propose best practice. English schools already have many of the tools that are needed. It is for the schools to use them to implement this good practice.

Ten Tips for Successful Teacher Evaluation

1. Schools should not rely on one single approach to teacher appraisal or evaluation. Instead they should consider using a mix of value-added or progress measures, classroom observations and pupil surveys. Ultimately the mix chosen should be at the discretion of the headteacher with knowledge of the strengths of each.

2. A clear system should be developed for teacher appraisal that is implemented fairly and consistently for all teachers.

3. External advice should be used, where possible, to assess the quality and standards of a school's system and to assure staff of its fairness and governors of its robustness

4. Staff sessions should be used to discuss the new system and help shape its effective implementation.

5. Staff involved in evaluation should be properly trained, and school leaders should ensure that they are working within the agreed standards for the school.

6. Good feedback is at the heart of successful evaluation, if it is to lead to improved teaching. School leaders should ensure that there is proper one-to-one discussion about the results of any evaluation.

7. While appraisal and evaluation should focus on classroom activity, teachers' contributions to extra-curricular activities, including sports, trips and clubs, should also be recognised.

8. Value-added or progress measures, rather than absolute test or exam results, should be the primary data used in evaluating performance, as they are the most objective and comparable assessment of a teacher's contribution. It is important that robust baseline data is used.

9. Developmental and evaluative classroom observations should be carried out separately, to promote honest feedback. It may make sense for peers to be involved in developmental observations but those for appraisal purposes being conducted by members of the school leadership team. There should be clear standards and protocols for observations, perhaps in a school handbook.

10. Pupil surveys should be clearly structured, be age appropriate, and should complement other measures.

Introduction

This report reviews three methods of teacher assessment available to headteachers and other school leaders in England and Wales. It is informed by the large and growing academic literature on both sides of the Atlantic and is supplemented with current examples from this country. The report concludes by recommending procedures to school leaders in light of the 2012 changes to the teacher appraisal regulations (DFE, 2012A).

The large impact a good teacher can make on a pupil's academic outcomes is now well established (Aaronson, Barrow, and Sander, 2007, Rivkin Hanushek and Kain, 2005 Rivkin et al. 2005 and Rockoff 2004). This is especially true for pupils from disadvantaged backgrounds: one year under the supervision of an excellent teacher is worth 1.5 years' of learning compared to 0.5 years with poorly performing teachers. In other words, for poor pupils the difference between an excellent and a bad teacher is a whole year's learning[1] (Hanushek, 1992).

Whilst many agree that teaching is the most important factor in schools for pupil achievement, the best way to assess who are the 'good' and the 'bad' teachers has yet to gain such wide agreement. This debate on how best to evaluate teachers is top of the education agenda both in the UK and the US. The Obama initiated Race to The Top programme provides additional funding for states that have implemented performance based standards reforms. This has lead researchers, practitioners and policy makers all to ask the same question, what is best method of measuring teacher effectiveness?

In the UK, the same question arises from the recent reforms to the national teacher standards and the revised appraisal regulations. From 1 September 2012, schools have had considerably more freedom to assess teachers in the way that they see fit, according to their own individual circumstances. Classroom observations no longer need to be pre-arranged or limited to a maximum of three hours over a year. The government has provided a model appraisal system, but has not provided any details on how the evaluations should be implemented, or where to look for this advice. With schools having the freedom to develop their own policies, the Sutton Trust is in a position to provide guidance to school leaders on methods of best practice drawn from empirical research. Furthermore, given the government's intention to accept the recommendations of the School Teachers' Review Body (STRB) to give more freedom to schools to set teachers' pay (STRB 2012), it is important for schools to be using reliable and informative metrics.

The consensus is that standard CV information, such as education and experience, has little to no predictive power on a person's teaching ability. A recent literature review found that, in 86% of the papers, teachers' education had no significant effect and in 66%, teacher experience was also insignificant. Another paper with very detailed information[2] on teachers' history found that they explained less than 8% of teacher quality (Aaronson et al., 2007).

Therefore, we look to the classroom as the place to assess teaching ability. This is not unique to teaching. In all professions, the ability and effort of a worker can only be fully measured in their workplace. Many questions remain hotly debated: what should be measured, how should it be measured, and how often? This report considers the three main methods of teacher evaluation; pupil test scores, classroom observations and pupil surveys. We highlight the main arguments for and against each whilst providing empirical

[1] Defining an excellent (or bad) teacher as a teacher one standard deviation better (or lower) than the average in terms of value-added test scores.
[2] This included gender, race, teaching experience, undergraduate university attended, advanced degrees, teacher certification and current tenure.

evidence which should be considered when deciding on an evaluation system.

England's decentralised evaluation system allows for discretion when making decisions based on these measures. A centralised decision making process that is prescriptive will undoubtedly lead to cases of misclassification, given the noise associated with these measures. Teacher evaluation metrics are not absolute and therefore they should only be used as indicators of performance. We must rely on the expertise of experienced teaching leaders to make informed decisions when appraising a teacher, taking all factors into account including those that effect achievement and the strengths of each measure.

This decentralisation also means teachers' activities outside of the classroom can be considered. Schools are complex working environments and a teachers' contribution to effective management and extra-curricular activities are also important.

All measures of teaching ability are imperfect, and cannot hope to capture all the complexity of the teaching profession. Each has its advantages and disadvantages. The weighting given to each method depends on the use to be made of the evaluations. Gains in pupil achievement are the single best predictor of future teaching ability, classroom observations provide valuable feedback in terms of teacher development, and pupil assessment can provide both insight where formalised testing is inappropriate and feedback on teaching style. A combination of all three measures will provide the most reliable and trusted outcome.

Gains in Test Scores

Pro: Highly Predictive
Con: Universal Applicability

Using the final grades pupils achieve, rather than their gains, is a bad measurement of teacher achievement. This is because the largest determinant of pupil achievement is their family background (Goldhader et al. 1999, Hoxby, 2001) and this is something which a teacher cannot change. It is for this reason that *gains* in test scores, or value-added measures, have become widely used for assessing teaching performance.

Value-added (VA) test scores use the gains in pupil test results whilst under the direction of the teacher, so they take into account initial pupil ability. In England, the school performance tables include a measure of value-added in primary schools and between the ages of 11 and 16 in secondary schools. There is also data showing the value-added for disadvantaged pupils, and for those regarded as low, middle and high attainers on the basis of previous tests.

However concerns still remain about the validity, stability and precision of such measures. This section takes an uncompromising look at how well the value-added metric measures up to these ideals, by asking whether value-added test score gains provide

- an unbiased measure of teacher quality
- a consistent measures over time; and
- an accurate reflection of teacher quality?

1. Is VA an unbiased measure of teacher quality?

For VA test scores to be an unbiased measure of teacher effectiveness we need to make four assumptions:

(i) Teachers are unaffected by their working environment;
(ii) Growth in test scores is a priori equal conditional on test scores (or pupil assignments to teachers are random once the prior test score is taken into account);
(iii) Test scales are invariant (that the percentage point gain is of equal value regardless of the baseline); and
(iv) Teachers are equally effective with all pupils.

The literature has tested each of these assumptions. While typically they are not found to hold true, they also have very little effect on the calculated value-added scores in practice.

(i) **Teachers are unaffected by their working environment:** Angrist and Lavy (1999) found that school facilities such as class size do have an effect on pupil learning, Case and Deaton (1999) also found that school administration and cooperation amongst the teachers improve pupil outcomes. This means that not all gains made by the pupil are due to the teacher, a teacher in a more effective school would have better value-added (VA) scores than the same teacher in a less effective school. However, this can be resolved by taking the school characteristics into account when calculating VA scores or, more simply, comparing teachers within a school.

(ii) **Growth in test scores is a priori equal conditional on test scores (or pupil assignments to teachers are random once the prior test score is taken into account):** If different pupils have different rates of growth in test scores and they are not randomly matched to teachers, this could bias the measures of teacher effectiveness (Rothstein 2009, Feng 2005). Consider a case where a teacher has a choice to teach privileged or non-privileged children. The

teacher should be indifferent between the two groups in terms of the value-added that they can provide. However, if the privileged students have more opportunities for additional learning outside of school, such as parental help with homework, tutoring or extracurricular activities, these pupils could have higher gains than the non-privileged pupils. This may hurt the poorer pupils as there would be incentives for teachers only to teach the more privileged. A similar situation arises with the ability setting of pupils within schools: teachers would prefer to teach the high ability students if they believed that their achievement growth rate would be higher than that of low ability students.

This is only a problem if teacher matching to pupils is not random - if teachers can choose their students within a school or consistently teach only one type of pupil. This could lead to systematic biases in the VA measurements. Kane and Staiger (2008) tested the extent to which this sorting affects VA test scores and found it only to have very minor effects. Furthermore, Koedel and Betts (2008) found that this is only a problem when focusing on single year measures from one class. Value-added scores of teachers who teach in many classrooms over many years remove nearly all biases that might result from pupil sorting. The exception is where there is a high degree of sorting and lack of mobility of teachers between classes.

Value-added methods that take into account the differing academic growth rates of pupils have been found to have an extremely high correlation to value-added measures that do not. Johnson et. al (2012) found a very high correlation (0.96-0.98) between VA measures that take into account pupil and peer characteristics using basic administrative data[3] and VA that don't; similarly, Ballou (2004) found negligible differences between the measures. Nevertheless, the few teachers who are systematically disadvantaged when pupil characteristics are not included are those who teach pupils from predominantly disadvantaged backgrounds. However, some researchers even argue that it is detrimental to disadvantaged pupils to allow for differential growth rates, as they will implicitly reduce the expectations of their teachers (Sanders et al. 2009).

(iii) **Test scales are invariant:** The value-added model assumes that test scales are invariant, that the gains made by pupils from improving the test score by five points are the same at all points across the score distribution. This implies that the gains of improving from 5% to 10% are equivalent to moving from 65% to 70% and 90% to 95%. Psychometricians who design tests do not make these claims and therefore we cannot assume that it is the case (Barlevy & Neal, 2012). A decentralised solution involves the headteacher and teacher agreeing on targets for each class or pupil. This would allow for the differences in scale and for individual circumstances to be taken into account.

(iv) **Teachers are equally effective with all pupils***:* Finally, value-added models assume that teachers are equally effective with all types of pupil. However, it has been found that teachers' impact on pupil learning is dependent on the pupil-teacher match (Dee, 2005; Carrell et al. 2010; Grönqvist and Vlachos, 2008). Therefore, it is also becoming important for school leaders to work efficiently matching pupils and teachers together optimally.

Despite these violations of the assumptions in practice, simple estimates of value-added have been found to be close to experimental estimates (Kane & Staiger, 2008). Taking into

[3] The state data for the student and their peers are reduced-price meals status, disability, ethnicity, and English as a second language, along with gender and the age of the student. Using more detailed district data the correlation between basic and conditional VA measures falls to 0.93-0.97, this includes information on previous attendance and suspensions, as well as prior peer achievement.

account the working environment of a teacher and pupil characteristics removes many systematic biases associated with a basic value-added measure. However, even without adjusting factors, a basic VA score is typically not qualitatively different from an ideal score. Other researchers have compared teacher value-added scores with headteachers' evaluations of teacher ability and found a close correlation (Rockoff et al. 2010, Rockoff and Speroni 2011). Emphasising a teacher's value-added can be a genuine reflection of their underlying ability, and isn't just about gaming of test scores.

2. Is VA a consistent measure over time?

We expect teacher quality not to vary to a great extent year on year. Therefore, when choosing a measure of teacher quality, we would also want stability in that measurement over time. However, it has been shown that single year value-added measures are not stable (Ballou 2005, Koedel and Betts 2007, Goldhaber and Hansen 2010, McCaffery et at. 2009). Koedel and Betts (2007) illustrate this instability by showing the annual movement in teachers VA ranking. If teachers were equally effective every year - and test scores were an accurate reflection of pupil ability - all teachers would stay in the same quintile from one year to the next, and the proportions on the main diagonal would all be 100% .This is clearly not the case: the majority of teachers move between quintiles each year. In their sample of 941 teachers, Koedel and Betts found that 13% of teachers in the top quintile in one year were in the bottom quintile the next.

So, when gains in test scores are being used as a measure of teacher effectiveness, this inconsistency needs to be taken into account. Measuring the change over a year's worth of teaching may not be representative of a teacher's ability. In this example, pupils in the first year could have done unusually well while the next set of pupils might perform unusually badly in the second year. The solution is not to judge teachers on a single year's VA measure; researchers have shown that when VA scores are averaged over a number of years they become much more stable and start to reflect the underlying impact of the teacher. McCaffrey et al. (2009) show that stability increases by 40–60% when aggregating data across two years and by a further 18–23% when a third year is included. Schochet and Chiang (2010) give another vivid example of the instability of single year estimates, compared to three year averages. They found that there is a 25% increase in the chance of an 'average' teacher being labelled to be exceptionally[4] bad (or good) by a single year measure; equally, the chance of being labelled average whilst being exceptional is also 25%.

Table 1 Persistence of Teacher Fixed Effects Estimates

		Teacher Quintile Rank				
		1 (Worst)	2	3	4	5 (Best)
Teacher Quintile Rank in Previous Year	1 (Worst)	30%	20%	19%	18%	13%
	2	23%	25%	13%	21%	18%
	3	18%	29%	25%	24%	13%
	4	15%	15%	25%	20%	23%
	5 (Best)	13%	17%	16%	19%	35%

Note: Based on 941 teachers, Koedel and Betts (2007)

[4] They define exceptional as over one standard deviation from the average, below the 17th percentile and above the 83rd.

3. Is VA an accurate reflection of teacher quality?

The final test of the usefulness of value-added to teacher evaluation is precision. Even if the VA estimates are not stable, that does not necessarily mean that they are inaccurate. They could be an accurate reflection of their changing impact. How accurately do value-added test scores reflect the ranking of teachers in a given year? McCaffery (2009) found that 30-60% of the variation in measured teacher performance is due to sampling error from "noise" in student test scores. Critics of value-added measures highlight the case of a pupil having an especially good day and getting high scores. It then becomes much harder for his or her later teachers to produce gains in test scores. Similarly if a child does unusually poorly on a test, his or her later teachers will find it easier to generate gains in test scores. Whilst this may be true in individual cases, the proportion of pupils having good or bad days should cancel each other out statistically, meaning that overall we will get a better picture of teacher impact. Much of this sampling error noise is driven by VA scores being generated from a relatively small number of pupils, so increasing the number of pupils (by averaging over years or classes) greatly reduces the likelihood of this error occurring.

Using single year gains in test scores would make it much more likely that a teacher would be misclassified. Therefore, test scores should only be used as indicative indicators of where a teacher is on the distribution of teacher effectiveness. Using three year averages of teacher value-added Ballou (2005) finds that 60% of maths teachers are significantly different from the mean, but single-year estimates only identify 30%. Value-added scores cannot reliably tell the difference between which teacher was at the 40th percentile versus the 60th - those just above or below average - however it could be used to identify those at the extremes, such as the top or bottom 5% of teachers.

The Measures of Effective Teaching (MET) project[5] in the United States, which is funded by the Bill and Melinda Gates Foundation as a partnership between 3,000 teacher volunteers and dozens of independent research teams, has formalised this by evaluating the risk of misclassification. It calculated that the probability of someone coming from the bottom 25% when VA test scores indicated that a teacher was in the bottom 25% was only 54%. However, using VA test scores only to identify the bottom 3% teachers reduces the risk of misclassification significantly. This time 80% of the group are in the bottom quartile.

Increasing the number of observations from which a value-added test score is generated would reduce this measurement error even further. As we have noted, this could be achieved by averaging over several years or classes.

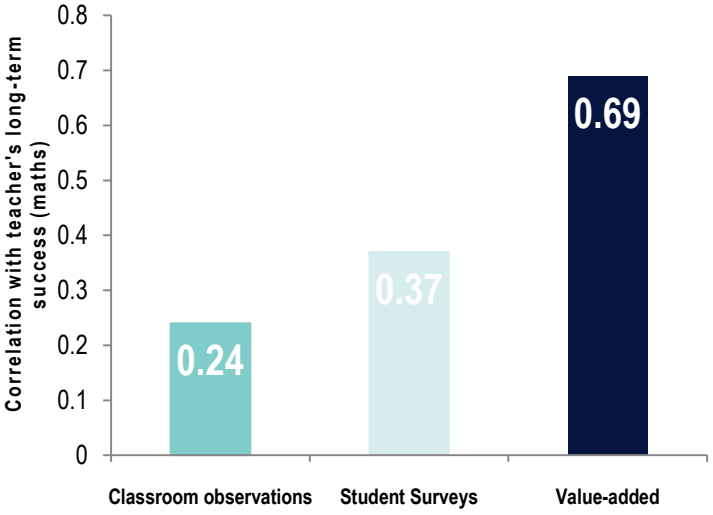
Another way to improve precision is to improve the underlying measure of quality. Many teachers complain that standardised testing does not test what they teach (or would like to teach). Pupil assessments would ideally measure intended outcomes of the course beyond test scores, such as knowledge, understanding and creativity. However, these characteristics are impossible to capture perfectly so instead we have a poor proxy. Standardised testing may give misleading results about the quality of teachers, so the argument goes, unless we place a low value on aspects that the test does not cover.

Much of this concern from America stems from their use of multiple-choice testing. It is easier to teach to these tests and they offer little opportunity for pupils to show their understanding of a subject. By contrast, exams set in England are more open, leaving it to the pupil to prove themselves, by showing how they reached particular conclusions and in their writing.

[5] www.metproject.org

Yet, even though the American tests need to be improved, it is still not true that teachers with larger gains are coaching students at the expense of other parts of the curriculum. The MET study found that more effective teachers not only caused students to perform better on state tests, but they also caused students to score higher on other, more cognitively challenging assessments in math and English (MET 2013) These students, also significantly, were more likely to enjoy class. Researchers have also looked at the long run effects of having an effective teacher measured in terms of VA. Chetty et al. (2011) tracked one million children from 4th grade to adulthood and find that those assigned to higher VA teachers are more successful later in life. These students are more likely to attend college, less likely to have children as teenagers, earn higher salaries, and live in better neighbourhoods.

Figure 1: Predictors of teacher's long run success



Source MET 2012

Value-added Conclusions and Applicability

As we have seen, there are problems with value-added measures of teacher effectiveness. They can potentially be biased, as all classes are different; they change year on year; and they cannot be relied upon to be accurate. However, despite these shortcomings, experimental estimates showed that the potential biases are very small in reality and that value-added test scores are by far the most predictive measure of teachers' long term success (MET 2012) (Figure 1). The differences in growth rates due to pupil or class characteristics have been shown to have little effect. Taking an average of test score gains over multiple years and classes removes other irrelevant factors and provides a cleaner measure of teacher effectiveness. Moreover, there is little evidence of coaching to a specific test: pupils who gain in tests used to calculate teacher value-added also improve in other tests. Most convincingly, value-added test scores also correlate strongly with headteacher assessment of ability (Rockoff et al. 2010, Rockoff and Speroni 2011) plus have the additional benefit that they are more objective

Gains in test scores are not a perfect way of measuring teacher impact. But they do provide a good starting point to establish which teachers are having the most impact on pupils. The foremost drawback of using test score gains as a method of teacher evaluation is a practical one: what test scores should be used?

England is in an excellent position to use gains in test scores as the National Curriculum sets out a framework for measuring pupil achievement. The Curriculum has a set of eight attainment levels which clearly set out what is expect from a pupil in terms of understanding and ability[6]. There are also sub-levels that allow for more detailed measures of pupils progress to be recorded. Objective assessments of pupil gains in achievement are available through nationally marked Key Stage (KS) test scores. However these are only conducted at the end of Key Stages 2, 4 and 5 (ages 11, 16 and 18) and so value-added measures between them could only be used to judge the effectiveness of a school as a whole, as multiple teachers and factors would of contributed to these gains in test scores. Furthermore, many teachers in secondary schools don't teach subjects

[6] There are current proposals to remove the system of levels, however to meet statutory requirements a new assessment system will have to be implemented. Therefore teacher assessment of pupil achievement would change as appropriate.

with baseline KS2 test scores[7] and so objective gains could not even be calculated.

Since the removal of externally marked KS3 assessment, many secondary schools in England have started to use independent testing agencies to assess pupil progress and attainment. In principle schools could use external agencies to test pupils in every year and in every subject though this would be expensive and disruptive. Secondary schools already conduct their own internal tests to measure the achievement levels of pupils. For these internal examinations to be informative on the levels of pupils' attainment it is crucial that these tests are calibrated to match with Key Stage achievement. This is a difficult task, but if it can be achieved, it will give schools a comparative measure of pupil progress across teachers.

There are various methods that can be used by schools to ensure that their tests are producing accurate measurements. The simplest way for secondary schools to perform internal moderation of exam results is to compare improvements of pupils to what is expected. If pupils in all Year 9 classes dramatically improve their test scores in History but have standard improvement patterns in Geography and English, a school may reasonably suspect that the History test was poorly calibrated. Furthermore, if the improvement in History is also radically different from the previous year's improvement, then we again may want to re-evaluate the test.

Schools with strong systems in place for analysing pupil data can gauge the validity of internal exams by examining how predictive they are of later externally marked test scores. If the internal test scores hold little relation to future test scores, the exam is either not well marked or not well-designed, so less weight should be applied to it. Of more importance to teacher evaluations are comparative rates of improvement: if the growth in pupils' scores between levels is high for one year but low in the next, this could mean that the grading was too lenient, or that one teacher was very effective, with a high value-added, and the other was not very effective. This is where the expertise of school leaders is important. They need to determine what has really happened.

Finally, the most direct method, which would assist such a verdict, is for schools to ask external experts to come in and train teachers in what measures of pupil attainment match with what's expected of pupils at a particular Key Stage level. This is potentially expensive but does have the advantage of schools receiving definitive confirmation of their marking schemes.

It is worth noting that these tests do not have to be precise. It isn't about differentiating between pupils at 51% and 53%, rather it is about correctly establishing at which Key Stage level or sub-level a pupil is performing (or the extent to which they are making expected progress for their age if levels are no longer used). Exact positioning does not make a difference to the pupil. Moreover, value-added measures themselves are not precise. For them to be effective, they just need results to be consistent and to be an unbiased measure over time and subjects.

In primary schools where there is little testing, teacher assessment of pupil achievement is commonly used. Critics of this method say that these measures are uninformative as teachers could inflate the grades of their pupils. Teacher assessments are valued on a basis of trust. Schools are small environments where there is little room for deception. It would quickly be known by teachers in subsequent years if pupils' levels had previously been misallocated. As teachers have to work with each other over many years, there would be a natural incentive to make honest reports of pupil achievement.

Work by Gibbons and Chevalier (2008) found that teacher assessments of KS3 did not consistently overestimate the ability of all pupils relative to

[7] Pupils are assessed in English, Maths at KS1, and in Science at KS2 and KS3. At KS4 all subjects are assessed.

externally marked exams. In fact they found that teachers tended to overly assume pupils were of average ability by underestimating the ability of high scoring pupils. However, they did tend to overestimate the ability of low performing pupils. A common concern held by many secondary school teachers is that end of primary school teacher assessments are inflated as they have no repercussions for the primary school. However, aggregate DFE figures (2012c) show that the proportion of pupils reaching Level 4 or above at the end of primary school by teacher assessment or externally marked examinations in maths are the same. To ensure no inflation of pupil grades, teacher should also be able to provide a portfolio of evidence of pupil achievement to match their assessment of the pupil if requested by their head of department or another school leader. And as these results are not about published league tables, there is no incentive for any collective misrepresentation.

Pupil achievement measures are already a core part of teachers' performance management in many schools today, including teacher assessments, internal test scores or national examinations. At the beginning of each school year, line managers agree their teachers' achievement targets for their pupils. These targets can take into account the current cohort of pupils along with their specific strengths and weaknesses. Having the teacher and line manager agree on targets that allow for diverse pupil growth rates succinctly deals with many of the potential problems with value-added measures.

Once all the test score information is available (generally at the beginning of the next school year), teachers are assessed if they have met these targets. Secondary schools may give more weighting to gains in test scores where both the previous year's and current test scores results were marked externally (years 11, 12 and 13), but with good internal test score data this can also be effectively done for every year.

There remains the issue that single year measures of teacher impact are unreliable: a teacher can make large gains in one year and very little in the next. Some parts of the English system already address this issue by making the progress onto the Upper Pay Scale[8] dependent on the previous two years' performance management targets and not just those in the previous year. In future, it is likely that schools will be expected to incorporate such measures into their overall approach to pay increases, as automatic increments are phased out. Additionally such annual variations mean that any value-added measures should not be used to distinguish between teachers just below or above the average, they can only be reliably used to identify the best and the worst teachers. Other career decisions should also only be made on the basis of multiple separate pieces of information.

Student test scores gains are a valuable metric for evaluating teacher impact. Despite the violations of the assumptions that underpin the model, the estimates are actually very close to experimental observations. They also closely reflect headteacher evaluations of teachers and are the most predictive of future achievement gains by other pupils. Because of large year on year variations, we should refrain from reading too much into single year measures and should be used as an indicator rather than an absolute measure. Gains in test scores can be used particularly effectively in English schools because the national key stage levels allow pupil achievement more easily to be compared across years and schools.

[8] The current pay system in English schools that use national pay agreements sees teachers rewarded on an incremental scale initially, with later increases dependent on performance. This may change with plans for a system based

Classroom Observations

Pro: Developmental Tool
Con: Unreliable

Teaching is unlike most professions, since the supervisor typically does not see the member of staff doing the job. The classroom separates teachers from other staff so that a teacher's ability is generally inferred from the behaviour and outcomes of pupils they teach. Classroom observations provide an opportunity for line managers and headteachers to see teachers in action. Here, they can assess style of teaching, pupil management and other aspects of teaching that cannot be obtained from other forms of teacher evaluation, such as value-added test scores. They provide an opportunity for teachers to receive constructive feedback on their teaching methods so that they improve over time. Personal evaluations also avoid classic arguments associated with test scores, such as 'teaching to the test', 'narrowing of the taught curriculum' and 'focusing on the marginal pupils' (Koretz, 2002).

But for all the benefits of an observation to be realised the observer must be properly prepared. This means they should have good training so that they know what to look for, can provide effective feedback and keep subjective opinions to a minimum. Effective training will also give teachers confidence in their evaluation, knowing that that it is meaningful and unbiased. There is relatively little economic research literature on classroom observations. But that which is available enables us to summarise the most important factors that make an observation successful.

Researchers in New York found that even a single observation of a trainee teacher was a significant predictor of later teacher quality (Rockoff and Speroni, 2011). Applicants to a teacher certification training programme were evaluated by professionals during an interview process which involved a mock teaching lesson and an interview. Even though these evaluations placed teachers on a very crude scale, with only five different categories, and had limited observational time, they were still found to be a strong predictor of future pupil test score gains. Those who were accepted onto the programme were observed during their first year of teacher training, and as the observation period grew longer, the reliability of these observational measures increased. Furthermore even when accounting for objective measures of teacher effectiveness, such as test score gains, these observational measures were still significant predictors of future performance. This implies that these subjective evaluations contained meaningful information about a teacher's effectiveness that is not captured in value-added measures.

However there are some important caveats to these results. The teachers being observed in this case were trainee teachers; the evaluators' job was to select the best. So, there was very little cultural or social pressure to be lenient in the observation process, unlike in other situations where teachers may be asked to evaluate their peers or work colleagues. Moreover, these observers were given training in evaluation and had explicit evaluation standards provided to them. One would therefore expect them to perform better than an average untrained teacher asked to perform the same task. Despite this training and professionalism, the researchers found that the implementation of these standards differed across observers. Some were a lot tougher than others. To the extent that it was inappropriate to make comparisons of assessments across observers without explicitly taking this into account, this highlights the importance of training the observers to ensure that their evaluations are informative and comparable. The implementation of any evaluation system should address this issue.

However, observations have not only been found to be useful for assessing trainee teachers. Jacob and Lefgren (2008) found that of classroom observation scores are strongly linked to gains in pupil test scores for established teachers. Teachers assessed through classroom observations to be one standard deviation better than the average would achieve the same gains in pupil test scores as a teacher who was one standard deviation better according to a value-added assessment. This research also found that when headteachers provided a teacher evaluation, they did not sufficiently take into account pupil characteristics and were overly influenced by absolute test scores. They tended to give teachers with poorer performing students a lower evaluation than a similar teacher with the same value-added, but higher absolute grades. Similarly research conducted by the University of Chicago Consortium on School Research (2012) found that staff members who were poorly trained in observations were more likely to rate teachers highly if that teacher had received high evaluation ratings in the past. This is one of the shortcomings of using teacher observations: they are inherently less objective than value-added measures.

Using headteacher opinions rather than formal classroom observations to appraise teachers has also been found to be effective. In a randomised intervention, Rockoff et. al (2010) found that headteachers' estimates of teacher effectiveness were accurate and become more so the longer they had worked together. This research also found that headteachers who were been given training in using student data started to include this information as part of their subjective evaluations. Headteachers gave test score gains more weighting when they were more precise and when they had spent less time with the teachers. Moreover, in schools where the information was provided, teachers of low ability were slightly more likely to leave; subsequently, objective pupil attainment data improved.

The advantage of such evaluations is that they are made over a long period of time, making it harder to 'game' in one-off observations, and they are not reliant on single year test scores. However, this is also the major disadvantage of informal appraisals. Without a set of standards against which teachers are assessed, a line manager will be open to bias. There is also no framework for teachers to improve their teaching.

Despite classroom observations being significantly correlated with teacher performance, they are still the least accurate measure of long-run teacher performance. The MET study (2012) compared the predictive ability of three measurement methods, observations, value-added scores and pupil surveys. They found that even when observers were highly trained, independent and calibrated each day, a single classroom observation was a far worse predictor of teacher success compared with value-added test scores or even pupil assessment. This is because an observation is only ever going to be a snapshot of what is going on in a classroom, whereas the other measures a come from a culmination of events over the academic year. Having multiple observations increased the reliability of observations and was further improved if the additional observations were conducted by different individuals even if they were for short time periods.

Classroom Observations: Conclusions and Applicability

Donaldson (2009) outlines the major factors that have limited the effectiveness of teacher evaluations in the past. These are classified into external and internal constraints. The external constraints comprise vague standards, restrictive labour agreements and a lack of time for evaluations. The internal constraints refer to the lack of training for evaluators, a school culture that discourages critical feedback and negative evaluation ratings, together with a lack of incentives for school leaders to evaluate accurately. These factors need to be considered when designing a teacher observation system. The majority of these concerns can be addressed by having a well-defined set of standards and well

trained observers; with these in place, the other gains will follow. Having well-trained observers with a clear framework keeps any subject biases to a minimum and ensures teachers have confidence in the evaluations.

Our conclusions for implementing an appraisal system draw on the results from the MET project. MET has spent the last two years evaluating five different methods of teacher evaluations[9] and provided advice for policymakers (MET 2012). They found that all the observational instruments produced very similar results, so policymakers should focus on their implementation rather than deciding which set of standards to use. The minimum requirements for good classroom observations, according to MET are:

1. **Choose an observation instrument that sets clear expectations:** Define a set of teaching competencies and providing specific examples at different performance levels
2. **Require observers to demonstrate accuracy before they rate teacher practice:** Teachers need to know observers will be fair and accurate.
3. **When high-stakes decisions are being made, multiple observations are necessary:** Averaging over multiple lessons reduces spurious evaluations.
4. **Track system-level reliability by double scoring some teachers with impartial observers:** To ensure reliability and keep teacher support, evaluations should be compared with those from external observers.
5. **Regularly verify that teachers with stronger observation scores also have stronger student achievement gains on average:** Even a great observation instrument can be implemented poorly.

So, how can these principles be applied to the English system? As part of the teacher appraisal system, schools are required to have in place a policy for classroom observation. The regulations surrounding teacher appraisal have been revised. The new regulations, which came into force in September 2012 (DFE 2012a), retain the key elements of the 2006 regulations but allow schools more freedom to design arrangements to suit their own individual circumstances. Restrictions on who does the appraisal, its primary purpose, advance warnings and total observation time have all been relaxed, giving school an opportunity to reform and improve their appraisal systems.

The key point is that although it doesn't matter greatly which particular rubric a school chooses to evaluate its teachers, it is very important that it has one. Any school without such a framework makes the task of assessor and assessed that much more difficult. In some cases, teachers are asked to assess their peers without being told what to assess, just that it needs to be done. In such circumstances, teachers could be providing unstructured and meaningless feedback to the classroom teacher.

There is no need for unstructured evaluation in England as there already are two national and well thought-through standards available - the national Teaching Standards and the Ofsted teaching standards. Mossbourne Community Academy, which is regarded as one of the most successful non-selective schools in England, combines the two standards to create a taxonomy of descriptions of teacher performance to be used in classroom observations. For the academy, this gives them the advantage of having clearly defined standards of what is expected of a teacher to be classified as Outstanding, Good, Requiring Improvement or Inadequate. These standards are aligned with the Ofsted categories for external inspections, against which the academy will be judged.

To ensure that these standards are being properly appraised, it is essential that those carrying out the appraisals are properly trained. This would involve

[9] Framework for Teaching (or FFT, developed by Charlotte Danielson of the Danielson Group), Classroom Assessment Scoring System (or CLASS , developed by Robert Pianta, Karen La Paro, and Bridget Hamre at the University of Virginia), Protocol for Language Arts Teaching Observations (or PLATO, developed by Pam Grossman at Stanford University), Mathematical Quality of Instruction (or MQI, developed by Heather Hill of Harvard University) UTeach Teacher Observation Protocol (or UTOP, developed by Michael Marder and Candace Walkington at the University of Texas-Austin).

setting out the agreed standards to all the teachers in the school. It may additionally require training days so that teachers know what they should be looking for in practice. This will give teachers the confidence to assign the appropriate levels. Unconfident teachers are less likely to award extreme marks (outstanding/inadequate) to deserving teachers. In the MET project, all observers were tested each morning against a calibration video. If they rated the teacher on the video significantly differently from their pre-rated level, that observer would not conduct any appraisals that day. This is obviously an extreme example, but it suggests how schools could train observers. It is worth noting that, even with these intensive methods, the MET project still found variation between its observers.

A classroom observation is only ever going to be a snapshot of what is going on in a classroom. However, having more observations gives evaluators more snapshots from which to generate a more complete picture. The teacher benefits because a single bad day is less likely to ruin an annual appraisal. The 2006 regulations limited the amount of time an average teacher could be assessed to three hours[10], but this restriction was removed in 2012.

Even with an increased number of observations, there is still a danger that some teachers would spend a lot of time preparing just for the observation class, making it unrepresentative. Evidence to support this was found amongst Chicago teachers, whose ratings were significantly lower in unscheduled observations (Chicago 2012). Schools having drop-in as well as pre-arranged observations can deal with this to some extent. For instance, Mossbourne Academy has two formal planned observations and two drop-in observations per teacher per academic year. In addition to assessing the class according to the Teacher Standards, teachers are also required to provide evidence of homework and marking in three sets of books. These are assessed on the quantity and quality of the marking and pupil feedback. Requiring marked homework is another way of extending the effective period of observation beyond a single class, making the evaluation more representative of the teacher's actual ability.

The best way for teachers to be confident of getting an unbiased and representative measure of their teacher effectiveness would be to have impartial observers conduct the appraisals. This happens to some extent currently with Ofsted inspections. Although they are not annual and they do not evaluate all teachers in the school, they can be used to calibrate internal measurements of teaching performance. But caution is required as such observations may be unrepresentative, as discussed earlier, and variations in classroom observations are much greater than value-added test scores for a teacher of a given ability. Schools could also pay for outside agencies to come in and observe classes. This would provide another opportunity to validate internal measures of ability and a chance to train teachers in effective observation methods.

The best way to obtain impartial measures of effectiveness without using outside agencies is to have a well-defined system in which staff members are accountable to the next level above them. Having a well-run management structure within schools provides a check on the observers to ensure that they are implementing the appraisal process correctly. It also means that when heads of department or year are given targets, they will have the incentives to provide the best feedback they can to their teachers. Of course, for that to happen they need a good appraisal system.

Finally, as we have seen in Figure 1, classroom observations are the least closely linked with long-run teacher performance, having approximately a third of the correlation of gains in test scores. This is because a classroom observation can only be a glimpse of the teaching process, whereas test

[10] Unless the teacher was at the capability stage of teacher appraisal process.

score gains are a culmination of the teacher's input over the course of a year. Because of this classroom observations should not have a large weighting in the formal assessment process.

The main potential benefit of classroom observation is that it allows for constructive feedback to the teacher, something which the other methods cannot provide. It has been shown that effective feedback has improved the long run effectiveness of teachers. Mid-career teachers in Cincinnati, Ohio, who took part in a local Teacher Evaluation System (TES), were evaluated in the classroom by three high-performing peers and their principals at four points in the school year, and they provided feedback. This was found to increase teacher value-added during that year of observation, but also in the years after the observation (Taylor & Tyler, 2011).

To promote uninhibited feedback from the observers, schools should separate the teacher appraisal and teacher development observations. This will give the observer and the observed teacher an opportunity for a free and frank discussion of the teacher's strengths and weaknesses without the concern of it being kept on permanent record. Use of distinct appraisal and development systems are in place in Arizona. Observational standards were first introduced as a developmental tool, which was eventually embraced by the teachers once it was established that the observations were useful and had no repercussions. Then teachers asked to be rated on these same standards that they had confidence in. To retain the advantages of both the development and formal assessment observations are separate and conducted by different observers, but both systems to use the same language and goals.

For classroom observations to achieve gains, it is important that the feedback given after the observation and as a part of the annual appraisal system is effective. There has been a lot of research in personnel literature on constructive feedback. The two most prominent approaches are 360-Feedback (Luthans and Peterson, 2003) and establishing of SMART targets (Doran, 1981). Both of these deserve their own research paper, but the common themes of each are specific measurable targets in an environment that encourages free discussion. The annual teacher appraisal process, used in conjunction with the national teacher standards, provides an ideal opportunity to put this into practice.

Pupil Surveys

Pro: Correlated
Con: Unclear determinants

Using pupil surveys to evaluate teachers has a long history in the research literature. The appeal of using pupils is that they are the ones who interact most with the teachers. Teachers can't 'game' the system as they can by preparing a class for an inspection, or pupils for a test. The surveys are based on the opinions of pupils built up over the school year, which advocates hope makes them harder to manipulate. There is evidence for their usefulness too: the MET (2012) project found that they correlated more with future pupil outcomes than classroom observations, even when the latter were conducted by highly trained independent observers. Despite such evidence, there is still much concern about what pupil surveys actually measure.

Historically, the major discussion concerned pupils' ability to rate their teacher, and to distinguish between how much they *like* a teacher and how *good* they think a teacher is (McKeachie, 1957). However, most research has shown that pupil surveys are correlated with pupil tests scores and value-added test scores. So now the debate is more concerned with the extent to which pupil evaluations merely reflect their grades rather than their actual learning? It may be a case of correlation rather than causation: do high ability pupils know that they are going to get good grades and so evaluate the teacher highly and is the same true in reverse for low ability pupils?

Whilst pupil surveys are still relatively rare in the English school system, the use of student evaluation of lecturers is now commonplace in the higher education sector (Becker and Watts, 1999). Therefore the majority of the research discussed in this section involves university students, though many of the findings will be applicable to the primary and secondary sectors.

Two recent pieces of research have cast further doubt on whether the correlation between higher value-added and pupil ratings in surveys represents a good teaching experience. Both studies use student survey data on lecturers' perceived ability and student test scores over a number of years to find that teachers who are given favourable student evaluations have high value-added in that year, but in subsequent years, the students of lecturers who had high pupil ratings did less well. Moreover lecturers who are associated with better subsequent performance receive poorer evaluations from their students. What is even more interesting is that the settings for these findings are very different: one was the US Air Force Academy (Carrell and West 2010) and the other a university in Italy (Braga, Paccagnella and Pellizzari, 2011)

An explanation for the common finding that teachers who are rated highly tend to have pupils who do well in their course but poorly in subsequent related courses was put forward by Braga et. al. (2011). Teachers can engage in real teaching or in teaching-to-the-test. The former requires higher student effort but generates real learning; the latter guarantees high grades for the current course but does not improve actual knowledge or future outcomes. Students prefer teachers that teach to the test, perhaps because they find it hard to tell the difference between the different methods, other than in the amount of effort they have to put in, or they simply have a preference for grades over learning. This is of concern for pupil evaluations, as the goal of good teaching should be learning that lasts as well as short-term grades.

Both the teaching to the test mechanism and high ability pupils rating teachers more highly depend upon the students' beliefs about their future test scores. In each case, one would expect survey

questions on achievement to be the best predictors of student test score gains. However the MET study found that students who described their learning environment as focused, engaging and demanding did even better. Even if students prefer a low effort learning environment, asking them whether the class is challenging still elicits important information about the teacher. This is the critical issue when discussing pupil surveys, what questions are asked? Questions relating to the classroom atmosphere are likely to be more indicative of teaching ability than a pupil's like or dislike for a teacher. Equally questions about factual aspects of the learning process could also prove informative: "How often are you set homework?" or "How often is your work marked?"

Pupil Surveys Conclusions and Applicability

Teachers ranked highly in pupil surveys have consistently been those who achieve the best grades from their pupils. However, the causal interpretation of some of these findings is being questioned. The most convincing work comes from the MET (2012) study and uses Cambridge Education's Tripod Project survey questions. These focus on the activities of the teacher rather than the pupil's feelings towards their teacher and are referred to as the 7Cs.

- Caring about students – "The teacher in this class encourages me to do my best."
- Captivating students - "This class keeps my attention – I don't get bored."
- Conferring with students - "My teacher gives us time to explain our ideas."
- Controlling behaviour - "Our class stays busy and doesn't waste time."
- Clarifying lessons - "When I am confused, my teacher knows how to help me understand."
- Challenging students - "My teacher wants us to use our thinking skills, not just memorize things."
- Consolidating knowledge - "My teacher takes the time to summarize what we learn each day."

These questions not only provide an overall appraisal of a teacher, but can also be used as a form of feedback to teachers to improve on their methods.

Although there has been a move towards giving pupils a greater say in English schools in recent years through programmes like Student Voice, pupil or student surveys are not common in England even if they are increasingly discussed amongst educationalists. The Welsh system allows for the opinions of pupils to be heard through the introduction of statutory school councils in 2005, but Welsh schools have yet to introduce pupil evaluation of teachers. However, a recent survey of Welsh teachers found that the majority of teachers who expressed a view had 'no problem with pupils rating their teaching' (TES, 2008). Ofsted has also produced student surveys though its questions have focused on the school as a whole rather than a particular teacher[11] and so would be inappropriate to use their data as part of an appraisal process.

Even if schools introduce surveys on teaching styles, there is another danger. Pupils would know that the surveys will reflect on the teachers and could provide answers to damage a particular teacher. It is for this reason that, if pupil surveys are used to evaluate teachers, we should be cautious in applying too much weight to them, even if they do correlate closely with test scores. These evaluations do provide some value: they can be useful to calibrate and feed into classroom observations and are also a good source of feedback to teachers about their methods, identifying what is and is not working amongst their pupils.

The most beneficial aspect of pupil surveys is that they can be used as an additional piece of evidence for line managers or teachers in the end-of-year appraisal process. As we have seen, test scores and classroom observations are a 'noisy'

[11] Pupils were asked to answer the following questions. 1 I enjoy school, 2 My school helps me to be healthy, 3 I feel safe when I am at school, 4 I learn a lot in lesson, 5 Behaviour is good at my school, 6 Adults in my school care about me, 7 Adults at school are interested in my views, 8 I know how well I am doing at school, 9 Adults explain to me how to improve my work, 10 My school helps me to get ready to move into my next class, 11 The headteacher and senior staff in my school do a good job. For KS2 pupils could agree or disagree, and for KS3/4 pupils could rate their level of agreement; Strongly Agree, Agree, Disagree or Strongly Disagree

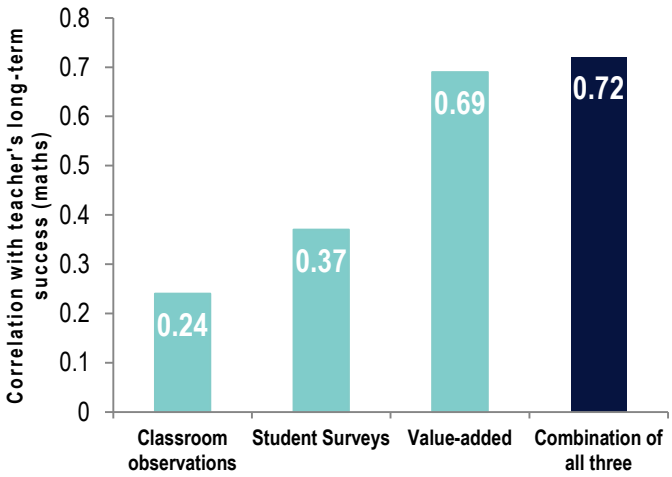
measure of teacher ability. If pupils performed poorly on test day, or an observed class did not go to plan, pupil surveys can help to assess the extent to which teaching throughout the year was consistent with previous years, and may suggest that the other measure was just an outlier. Given the unpredictable nature of education, all pieces of information are useful in reducing this noise and obtaining estimates closer to the truth.

Combining Measurements

We have seen that classroom observations, pupil surveys and value-added measures are all informative in identifying effective teachers. We have also seen evidence that each provides information that the other does not. In New York researchers found that classroom observations were still a significant determinant of future pupil gains even when teacher value-added was taken into account (Rockoff and Speroni, 2011). This leads us to ask the question, what combination of measures is the best at predicting teacher effectiveness?

The MET project addressed precisely this question. In the first phase (MET 2012) they concluded that combining all three measures was more correlated with long-run teacher success than any single measure (Figure 2). The second phase of the report examined which combinations provided the best measurements for gains in future test scores, other higher order thinking test scores and reliability[12] (MET, 2013). A system that applied a lot of weight to previous gains in test scores were the best at predicting teachers who would produce future gains in test scores. However, these systems were also the least reliable, reflecting that single year gains in test score measures have considerable variation. On the other hand a system that gave large weighting to classroom observations had the lowest correlation with test score gains. The systems that performed well in all three categories, including higher order thinking, were ones that were comprised of 33%-50% value-added measures with the remainder equally split between more stable measures such as student surveys and observations.

Figure 2: Predictors of teacher's long run success



Source MET 2012

Combining Measurements Conclusions and Applicability

With the introduction of the Race To The Top framework in the US, the urge to identify who are the most effective teachers has increased. The most adopted method for doing this is to use multiple different measures as it is seen as most fair and valid. These are typically combined through a state specific weighting system into a single index of teacher effectiveness which are then used to inform decision making. As previously seen using multiple measures is beneficial as it makes the final score more accurate and reliable. It also takes much of the decision making out of the school administrators hands which removes potentials for bias. Furthermore as these measurements take into account different aspects of teaching it will restrain teachers from focusing their and their classes attention on just one specific outcome. For example an over weighting on test scores could lead to increases in teaching to the test at the expense of pupil creativity or enjoyment of the subject.

The shortcoming of this system is that because the weighting system between the measures is decided centrally and typically the score is computed centrally that it makes the system highly prescriptive. It doesn't allow for on the ground

[12] Reliability was defined as year to year stability of teachers results.

experience to enter into the decision process. This is problematic given the imprecise nature of the metrics that make up the index. However, this can be remedied by only taking action when teachers are only seen at the extreme ends of the distribution over a period of years.

In the UK the teacher evaluation system is decentralised to the school level. This allows for more flexibility, so that factors in and out of the classroom can be taken into account. A headteacher will know if a particular class is abnormally disruptive, or that there were outside problems on the day of the observation. This means that headteachers have discretion in how they weight each measurement, which would hopefully reduce the chances of misclassification. Local decision making also allows for a wide range of potential teacher activity outside of the classroom to contribute, such as effective management and extra-curricular activities.

This puts a considerable amount of trust in the expertise of experienced teachers to make the right decisions. Therefore it is important that headteachers are accountable for their actions. This could be achieved through a range of channels from governor reviews, to the rewards for high performing teachers coming out of the school budget and would ultimately be seen in the demand for school places. For headteachers to make informed decisions when evaluating teachers it is critical for them to be aware of the strengths and weakness of each of the measures so that they can be taken into account.

Conclusions

This report reviews three methods of teacher assessment available to headteachers in England and Wales. It is informed by the large and growing academic literature on both sides of the Atlantic and is supplemented with contemporary examples from England. Each method has weaknesses, but each has its appropriate use within a comprehensive teacher evaluation system.

Gains in test scores for teacher performance: Gains in pupil test scores are the best available metric to measure teacher performance. Improvements in student attainment are an imperfect measure, but they are a starting point. The main advantage of this measure is its objectivity and despite its shortcomings is by far the most reliable of the three measures in predicting a teacher's future performance. Schools in England are ideally placed to use this measure as the Key Stage achievement levels provide common datasets over time.

Classroom observation for teacher development: Even when conducted by well-trained independent evaluators classroom observations are the least predictive method of assessing teacher effectiveness. However being observed does allow for an unrivalled opportunity to provide constructive feedback to teachers. To promote honesty in the feedback developmental and evaluative observations should be carried out separately. Observations are common in schools in England today but, for them to be most effective, clear standards must be established. Again, schools in England have standardised measures of teacher performance that can be used to this effect.

Pupil surveys for corroborating measures: Whilst pupil surveys are open to accusations of misreporting by pupils, it has been found that they do contain information on the effectiveness of the teacher. Whilst student surveys are not as predictive as test score gains, nor do they provide as much effective feedback as peer observation, they do provide a middle ground against which gains in test scores and classroom observations can be calibrated.

Decentralising the evaluation of teachers to schools allows for more flexibility, so that factors in and out of the classroom can be taken into account. Using a centralised rules system to determine the best and worst teachers will undoubtedly lead to cases of misclassification, given the noise associated with these measures. We rely on the expertise of experienced teachers to take into account such factors when appraising a teacher. Decentralisation also allows for the wide range of potential activity teachers provide outside of the classroom such as contributing to effective management and extra-curricular activities. No measurement is perfect, as all measurements suffer from noise and can be driven by outliers. However, with knowledge of their shortcomings, we put forward what the evidence shows to be best practice. English schools already have many of the tools that are needed. It is for the schools in a system where they now have greater freedom to develop best practice. Combining each of these measures to produce a composite score of teacher effectiveness has been shown to be the most correlated with the long term success of teachers (Figure 2). Each measure adds different information to the overall assessment, and reduces variance. Even though gains in pupil test scores is the most reliable measure of teaching ability, classroom observations and pupil test scores are excellent sources of feedback that can be used to improve the teaching.

References

Aaronson, D., L. Barrow, and W. Sander (2007) "Teachers and Student Achievement in the Chicago Public High Schools." Journal of Labor Economics 25 (1): 95–135

Angrist, J.D., and V. Lavy (1999) "Using Maimonides' Rule to Estimate the Effect of Class Size on Student Achievement." Quarterly Journal of Economics 114(2): 533-575

Ballou, D. (2009) "Test scaling and value-added measurement" *Education Finance and Policy* 4 (4): 351–83.

Barlevy G. and D. Neal (2012) "Pay for Percentile" American Economic Review, American Economic Association, vol. 102(5): 1805-31, August.

Becker, W.E., and M. Watts (1999) "How departments of economics should evaluate teaching," American Economic Review (Papers and Proceedings), 89(2): 344–349

Case, A., and A. Deaton (1999) "School inputs and educational outcomes in South Africa." Quarterly Journal of Economics 114(3): F1047-F84

Carrell, S.E., and J. E. West (2010) "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors", *Journal of Political Economy* 118(3): 409-432

Carrell, S.E., M.E. Page, J. E. West (2010) "Sex and Science: How Professor Gender Perpetuates the Gender Gap" Quarterly Journal of Economics 125 (3)

Chetty, R., J.N. Friedman and J.E. Rockoff (2011) "The Long-term Impacts of Teachers: Teachers Value-Added and Student Outcomes in Adulthood" NBER Working Paper No. 17699, December 2011

Chevalier, A., and S. Gibbons (2008) "Assessment and age 16+ education participation, *Research Papers in Education , 2*3 (2) 113-123, June 2008 Working paper.

Dee, T. S. (2005) "A Teacher Like Me: Does Race, Ethnicity, or Gender Matter?" American Economic Review 95 (2): 158–16

Doran, G. T. (1981). There's a S.M.A.R.T. way to write management's goals and objectives. Management Review, Volume 70, Issue 11(AMA FORUM), pp. 35-36

Department for Education (2012a) "Teacher appraisal and capability: A model policy for schools", May 2012, https://www.education.gov.uk/publications/eOrderingDownload/Teacher%20appraisal%20and%20capability%20-%20model%20policy.pdf

Department for Education (2012b) "Teachers' Standards" May 2012, https://www.education.gov.uk/publications/eOrderingDownload/teachers%20standards.pdf

Department for Education (2012c) "National Curriculum Assessments at Keys Stage 2" September 2012, http://www.education.gov.uk/researchandstatistics/datasets/a00213778/national-curriculum-assessments-ks2-england

Department for Education (2011) "Post-Threshold, Excellent Teacher and Advanced Skills Teacher Standards", December 2011 http://media.education.gov.uk/assets/files/pdf/s/independent%20review%20of%20teachers%20standards%20%20%20second%20report.pdf

Donaldson, M.L. (2009) "So Long, Lake Wobegon? Using Teacher Evaluation to Raise Teacher Quality" Centre for American Progress

Feng, L. (2005) "Hire today, gone tomorrow: The determinants of attrition among public school teachers", MPRA Paper No. 589, University Library of Munich.

Goldhader D.D., D.J. Brewer, D. J. Anderson (1999) "A Three-way Error Components Analysis of Educational Productivity" Education Economics Vol. 7(3)

Goldhaber, D. and M. Hansen (2009) "Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions", Center on Reinventing Public Education Working Paper #2009_2.

Goodman. S, and L. Turner (2010) "Teacher Incentive Pay and Educational Outcomes: Evidence from the NYC Bonus Program", Program on Education Policy and Governance Working Papers Series, PEPG 10-07

Grönqvist, E. and Vlachos, J. (2008) "One Size Fits All? The Effects of Teacher Cognitive and Non-Cognitive Abilities on Student Achievement" CEPR Discussion Paper No. DP7086.

Hanushek E.A. (1992) "The trade-off between child quantity and quality." Journal of Political Economy 100, no.1 (February):84-117

Harris, D. N., and T. R. Sass (2009) “What Makes for a Good Teacher and Who Can Tell?” Calder Center Working Paper 30.

Holtzapple, E. (2003) “Criterion-Related Validity Evidence for a Standards-Based Teacher Evaluation System.” Journal of Personnel Evaluation in Education, 17(3): 207-219

Hoxby, C. M. (2001) "If Families Matter Most, Where Do Schools Come In?" in T. Moe, ed. A Primer on American Schools. Stanford: Hoover Institution Press.

Johnson, M, S. Lipscomb, B. Gill (2012) “Sensitivity of Teacher Value-Added Estimates to Student and Peer Control Variables” Manuscript. Mathematica Policy Research, Cambridge, MA USA

Kane, T.J., D.O Staiger (2008) "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation" NBER working paper No. 14607, December 2008

Koedel,C., and Betts R.J. (2007) “Re-Examining the Role of Teacher Quality In the Educational Production Function”, Working Papers 0708, Department of Economics, University of Missouri

Koedel,C., and Betts R.J. (2008) “ Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique”, Working Papers 0708, Department of Economics, University of Missouri

Koretz, D.M. (2002) “Limitations in the Use of Achievement Tests as Measures of Educators’ Productivity”, Journal of Human Resources 37(4): 752-777

Luthans, F. and S. J. Peterson (2003) "360-degree feedback with systematic coaching: Empirical analysis suggests a winning combination." Human Resource Management, 42(3): 243-256.

McCaffrey, D.F, T. Sass, J.R Lockwood and K. Mihaly (2009) “The Inter-Temporal Variability of teacher effects estimates” Education Finance and Policy 4 (4): 572–606

McKeachie, W.J. (1957) “Student Ratings of Faculty: A Research Review“ Improving College and University Teaching Vol. 5, No. 1 (Winter, 1957), pp. 4-8

MET (2012) “Gathering Feedback for Teaching Combining High-Quality Observations with Student Surveys and Achievement Gains”, Measures of Effective Teaching (MET), Bill & Melinda Gates Foundation

MET (2013) “Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment”, Measures of Effective Teaching (MET), Bill & Melinda Gates Foundation

Rivkin, S.G., Hanushek,E.A, and Kain,J.F.(2005) “Teachers, schools and academic achievement”, Econometrica, 73(2): 415–458

Rockoff,J.E (2004) “The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data” The American Economic Review , Vol. 94, No. 2

Rockoff, J.E., and C.Speroni. (2010). "Subjective and Objective Evaluations of Teacher Effectiveness." *American Economic Review*, 100(2): 261–66

Rockoff, J.E., D.O.Staiger, Kane,T.J., E.S Taylor, (2010) “Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools”, NBER Working Paper No. 16240, July 2010

Rothstein, J.M. (2009) Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy* 4 (4): 537–71.

Sanders, W. L., Wright, S. P., Rivers, J. C., and J. G. Leandro. “A Response to Criticisms of SAS® EVAAS®.” SAS® White Paper, 2009.

Schacter, J, and Y. M. Thum (2004) “Paying for High- and Low- Quality Teaching.” Economics of Education Review, 23(4): 411-440.

Schochet, P. Z. and H.S. Chiang, (2010) “Error Rates in Measuring Teacher and School Performance Based on Student Test Score Gains.” Washington: National Center for Education Evaluation and Regional Assistance

School Teachers Pay Review Body (2012) “School Teachers' Review Body: 21st report – 2012” http://www.education.gov.uk/schools/careers/payandpensions/a00203870/strb-remit-21st-report

Slater, H.,
Davis, N., and Burgess S. (2009) “Do teachers matter? Measuring the variation in teacher effectiveness in England” CMPO Working Paper No. 09/212

Springer, M.G., Ballou, D., Hamilton, L., Le, V., Lockwood, J.R., McCafrey, D., Pepper, M., and Stecher, B. “Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching,” Nashville, TN: National Center on Performance Incentives at Vanderbilt University, 2010

Sutton Trust (2011) “Improving the impact of teachers on pupil achievement in the UK” Sutton Trust

Taylor, E.S., Tyler, J. H. (2011) “The Effect of Evaluation on Performance: Evidence from Longitudinal Student Achievement Data of Mid-Career Teachers”, NBER working paper #16877

TES (2008) “Happy to be rated” http://www.tes.co.uk/article.aspx?storycode=2593678, TES Newspaper, 14 March

Vigdor, Jacob L. (2009) “Teacher Salary Bonuses in North Carolina,” Performance Incentives: Their Growing Impact on American K-12 Education, edited by Matthew Springer, Brookings, 2009

Wragg E., G Haynes, C. Wragg and R. Chamberlin (2001) “Performance Related Pay: The Views and Experiences of 1000 Primary and Secondary Head Teachers” University of Exeter, School of Education, Teachers' Incentives Pay Project Occasional Paper 1