CSE CENTRE FOR STRATEGIC EDUCATION
Leading educational thinking and practice

SEPTEMBER 2014

occasional paper 137

The formative evaluation of teaching performance

Dylan Wiliam

Dr Dylan Wiliam is Emeritus Professor of Educational Assessment at the Institute of Education, University of London where, from 2006 to 2010, he was its Deputy Director. An earlier version of this paper was presented at the International Conference on 'Educational Evaluation for Teaching Performance' organised by the Instituto Nacional para la Evaluación de la Educación, Mexico City, Mexico, 4–6 December 2013.

Introduction

The argument of this paper rests on four propositions. First, higher educational achievement is necessary both for individuals and for society. Second, higher educational achievement requires increased teacher quality. Third, increased teacher quality requires investing in the teachers already working in our schools. Fourth, that investment needs to take a radically different form from the professional development that teachers have typically received.

In Section 1 of the paper I briefly review research on the benefits of education for individuals and society. Sections 2 and 3 show that the quality of teaching in schools is one of the most important factors in determining how much students learn in school. Sections 4 and 5 show that the available evidence suggests that improving the quality of entrants into the teaching profession and removing ineffective teachers are inherently difficult, and unlikely to secure the kinds of improvement in teaching quality that are needed, leading to the central premise of Section 6, which is that the key to improvement of educational outcomes is investment in teachers already working in our schools. Readers who are willing to accept this premise may therefore choose to skip straight to Section 7, which shows that feedback, while potentially powerful, is often counter-productive, and requires careful attention to the context in which it is given, or to the exploration of this crucial element in Section 8. Section 9 proposes a model of formative evaluation that identifies five key strategies of formative assessment, and Section 10 provides an outline of how we might begin to think about the validity of formative evaluations of teacher performance. Section 11 concludes the paper with some principles for the effective implementation of formative evaluation of teacher performance.

© 2014 Centre for Strategic Education **Occasional Paper No. 137**, September 2014

ISSN 1838-8566
ISBN 978-1-921823-60-2

Centre for Strategic Education
(CSE) is the business name for IARTV
ABN 33 004 055 556
Mercer House 82 Jolimont Street
East Melbourne Victoria 3002
Phone +61 3 9654 1200
Fax +61 3 9650 5396
Email info@cse.edu.au

www.cse.edu.au

137

Section 1: Education matters, for individuals and society

Education matters for both individuals and society. For individuals, the benefits include increased lifetime earnings, better health and longer life. In addition, more educated individuals are less likely to be teenage parents, or to be involved with the criminal justice system (Wiliam, 2011a). For society the benefits are, if anything, even greater. More educated people are more tolerant and more likely to be involved in a range of prosocial activities (Feinstein et al, 2008), and they also make a greater contribution to economic growth. For example, Hanushek and Woessmann (2010) estimated that raising the scores achieved by Australian 15-year-olds on the PISA tests administered by the OECD by 25 points (the improvement made by Poland in a decade) would have a net present value of US$2.5 trillion. Perhaps even more surprisingly, just having all students in Australia achieving a score of 400 on PISA – the OECD's estimate of the skill level needed to function effectively in a complex society – would have a net present value of US$2.1 trillion for the Australian economy.

although having each school being as good as the best would of course improve education in Australia, it would not produce the kinds of improvements that are needed.

Section 2: Teaching quality is the crucial variable

In many jurisdictions, (eg, the United States, England), the emphasis has been on improving the quality of schools. This has intuitive appeal – all parents want their children to attend good schools – but what is surprising is that in most countries, as long as you go to school, it does not matter very much which school you go to. In terms of the OECD's PISA tests, around 7 per cent of the variation in the scores achieved by students in Australia can be attributed to the school attended by the student (McGaw, 2008). The remainder is made up of within-school variation (78 per cent), between-school variation explained by the social background of the school (7 per cent) and between-school variation explained by the social background of the students attending the school (8 per cent).

Of course, this should not be taken as implying that there are no bad schools. There undoubtedly are bad schools, and it may well be that for some schools, the best option may be to reconstitute the school by replacing the leaders, and possibly even the majority of the teachers (Bryk et al, 2010). What I think is important to understand is that the differences in progress made by students in different schools in Australia are small, and much smaller than the kinds of improvements needed to meet the challenges of the 21st century. In other words, although having each school being as good as the best would of course improve education in Australia, it would not produce the kinds of improvements that are needed.

The reason that the differences between schools are so small (once we take into account the differences in social backgrounds of the students attending those schools) is that the main determinant of the progress made by students in school is, perhaps obviously, the quality of instruction they receive. What is less obvious is that the quality of instruction received by students as they progress through a school is very variable. In all schools there is some great teaching, and some that is not very good, but, as students progress through school, the average quality of instruction received by students over their school careers does not vary greatly from school to school.

Section 3: Teaching quality is not the same as teacher quality

The quality of teaching depends on a number of variables, such as the amount of time teachers have to prepare instruction, the kinds of resources available, the number of students in a class, the skill of the teacher, and so on. In some systems (eg, Japan, Finland), the number of hours that teachers will spend actually teaching students is below 700 hours per year, while in others (eg, United States, Chile), it is well over 1000 (OECD, 2013a, p 396). It is important to realise, therefore, that teaching quality is much more than teacher quality. That said, it does seem that the quality of teachers in a system is a critical variable. If we allocate teachers into three equal-sized groups – 'below average', 'average' and 'above average' – then students taught by an above average teacher make 50 per cent more, and those taught by a below average teacher make 50 per cent less progress than students taught by average teachers (Hanushek, 2011). The most effective teachers are therefore at least 3 times as effective as the least effective. In fact, the differences in teacher quality may be even greater than this, because children do make progress, especially in language development, simply as a result of maturation. Indeed, one study (Fitzpatrick, Grissmer and Hastedt, 2011) estimated that one-third of the progress made by seven-year-olds was a result of maturation, so that it is likely that the most effective teachers are at least five times as effective as the least effective. Moreover, in both elementary (Hamre and Pianta, 2005) and secondary schools (Slater, Davies, and Burgess, 2008) it has been found that the best teachers benefit lower achievers more, so increased teacher quality closes the achievement gap. This does not mean that we should focus only on teacher quality. Ensuring that teachers have the resources they need to do their job is important. They need time, material resources, and the support of leaders and colleagues to do their best work. However, the magnitude of the differences between teachers in their effects on student learning means that improving the quality of teachers must be a priority for any education system.

In many countries, this has resulted in attempts to improve teacher quality by replacing existing teachers with better ones, through a combination of improving the quality of entrants into the profession (Barber and Mourshed, 2007) and removing ineffective teachers (Hanushek, 2010), approaches which are discussed in the following two sections.

Section 4: Predicting who will be good teachers is almost impossible

The fact that teacher quality is the most important ingredient of an effective education system does not, of itself, indicate the kinds of policies that can secure high-quality teachers.

Raising the bar for entry into the teaching profession looks like an attractive policy option, especially since high-performing jurisdictions tend to recruit teachers from the highest one-third of college graduates, and recent data from the OECD adult skills survey suggests that

the best teachers benefit lower achievers more, so increased teacher quality closes the achievement gap.

the correlation between the numeracy scores of teachers in a country and that country's score on PISA is around 0.5 (OECD, 2013b). In some high-performing countries, such as Finland and Singapore, there are ten to twenty qualified applicants for every place on teacher training programs, so that in addition to high-level academic qualifications, applicants need good communication skills and the necessary personal qualities to be effective practitioners.

For countries in this fortunate position, it might seem as if almost nothing else matters; if you are lucky to have the smartest people in your country wanting to be teachers, then you can get many of the other pieces of the puzzle wrong and still have a high-performing education system. However, it is worth noting that highly selective admission to teacher education does not guarantee a good education system. In the Republic of Ireland, admission to teacher education remains, as it has been for many years, highly selective (IRPSITEPI, 2012), and yet the country's performance on PISA in 2009 was indistinguishable from that in the United Kingdom. Conversely, in Shanghai, teachers typically do not have high educational qualifications, but are given extremely high-quality training both before and during their careers. Selecting teachers from the academically most able would therefore appear to be neither a necessary nor a sufficient condition for securing high teacher quality.

it seems to be extraordinarily difficult to identify who will be good teachers until they are in front of a class

Indeed, it seems to be extraordinarily difficult to identify who will be good teachers until they are in front of a class (see Gladwell, 2008, for a summary of the argument and the evidence). There is some evidence that students taught by teachers with higher academic achievement or IQs do make more progress (Slater et al, 2008; Hanushek, 1971) but the correlation is modest, and other studies (eg, Harris and Sass, 2007) find effectively no relationship between student achievement and the preservice education or academic credentials of the teacher.

There is evidence that well-structured interviews have some utility (see for example, Dobbie, 2012) but the correlation is again modest, and thus there is a real risk of rejecting those who would be very good teachers and accepting those who will not. More importantly, improving teacher quality by raising the bar for entry into the profession takes too long. If the bar for entry into the profession were raised, it would take at least 30 years before the last of those who entered the profession before the bar was raised left teaching.

One approach to raising teacher quality that is particularly popular at the moment is through elite programs, such as Teach for America and Teach First in England, in which high-achieving graduates undertake to teach in socioeconomically disadvantaged areas for a specified period of time. Evaluations of these schemes have not yet shown clear evidence that they are superior to traditional routes into teaching, even though they tend to be more expensive than traditional teacher education programs. Such schemes may raise the status of teaching as being a job that is worthy of the highest achievers, but the very design of such programs, together with the fact that they are explicitly 'elite' programs, means that the proportion of teachers 'in post' coming in through these routes is unlikely to exceed 1 per cent of the teaching force, even under the most optimistic assumptions.

Section 5: Evaluating teacher quality is inherently difficult

If changing the quality of entrants into the profession is difficult, then an obvious alternative would appear to be to remove ineffective teachers, but identifying ineffective teachers is rather more difficult than it might first appear. Observation protocols, such as the Framework for Teaching developed by Charlotte Danielson (1996), do 'work' in that students taught by teachers who are rated more highly on the framework do learn more, but these frameworks are unable to identify all, or even most, aspects of effective teaching.

For example, Sartain et al (2011) found a clear positive relationship between teacher ratings on the Framework for Teaching and

the amount of progress made by their students. Students taught by teachers who were rated as 'distinguished' (the highest level on the 'Framework') made approximately 30 per cent more progress than students taught by teachers rated as 'unsatisfactory' (the lowest rating). This is an important finding. Many previous attempts have failed to establish any clear link between observable teacher behaviours and the progress made by their students, so the fact that we can now (for teachers in the US at least) train people to rate teachers in ways that yield accurate ratings of teacher quality is an important step forward. However, as was noted above, the best teachers are at least 300 per cent more productive than the least effective (since the best teachers produce 18 months progress in the same time that the least effective produce 6 months progress). This suggests that the Framework for Teaching captures only around 10 per cent of teacher quality. A number of studies have concluded that, because teacher quality is not the same as teaching quality (in other words, because teaching performance is inherently variable), a large number of independent observations is needed to produce estimates of teacher quality that are sufficiently reliable to support high-stakes decisions such as termination of employment. For example, Hill et al (2012) found that just to get the reliability of teacher observations up to 0.90 (arguably a lower bound for high-stakes decisions) a teacher would need to be observed teaching six different classes and each lesson would need to be rated by five independent observers.

Therefore, while observation frameworks such as the Framework for Teaching do reliably identify aspects of teacher quality, when such frameworks are used for teacher evaluation purposes, because they account for so little of the variance in teacher quality, there is a real danger that teachers might become less effective even though they raise their ratings on the framework.

Given the inherent unreliability of teacher observations, a number of writers have argued for supplementing evidence from teacher observations with other sources of information, such as student evaluations of teaching and measures of the academic progress made by students (Kane and Staiger, 2012).

Measures of academic progress (often called 'value-added' measures) do appear reliably to identify different aspects of teacher quality from observations (Rockoff and Speroni, 2011), but estimating the value added by a teacher is extraordinarily difficult, even when we take into account prior student achievement, since most assessments under-represent the important outcomes of education. For example, good teachers continue to benefit students for at least two years after they stop teaching them (Rothstein, 2010). In other words, good teachers make the teachers who teach their students in future years look better than they really are. A second problem with value-added models is that differences in the statistical assumptions made in the modelling process show large variations in the ratings that are produced of teachers. For example Goldhaber, Goldschmidt and Tseng (2013) found that 9 per cent of the teachers who were rated in the top 20 per cent for value-added in one model (a random-student-effects model) were rated in the bottom 20 per cent with a traditional value-added model. Given also that the ratings of a teacher's value-added vary considerably from year to year (McCaffrey et al, 2008), value-added measures of student growth are problematic as indicators of teacher quality.

good teachers make the teachers who teach their students in future years look better than they really are.

The main conclusion to be drawn from all the attempts to assess teacher quality is that because all the measures are unreliable, we have to make a judgement about the 'burden of proof' required to identify inadequate teachers.

If we set the burden too high, then too few low-quality teachers are identified. For example, Winters and Cowen (2013) found in a study of reading teachers in Florida that if the criterion for removal of teachers was set as being in the lowest 5 per cent of value-added for two consecutive years then only 1 teacher in every 500 would be identified for removal. Of course the burden of proof could be relaxed, leading to more teachers being removed, but this would also lead to a greater number of highly effective teachers also being removed. In this context, it is also worth noting that Atteberry et al (2013) found that teachers who were identified as highly effective in their first year did not improve (as measured by value-added) over the first five years of their teaching careers, while those who rated as least effective in their first year of teaching improved steadily.

improving the performance of serving teachers will need to be the major component of every country's strategy to improve teacher quality.

The cumulative effect of all of the policy prescriptions listed above, even if implemented effectively and faithfully, would be small, and would take many years to materialise. For this reason, improving the performance of serving teachers will need to be the major component of every country's strategy to improve teacher quality.

Section 6: Professional development is the key to teacher quality

The foregoing may seem like a counsel of despair but the research on expertise in a number of different areas suggests that the teachers already in our schools could be much more effective than they are currently. There is now increasing evidence that measures of general ability are good predictors of how well someone does something only in the beginning stages. For example, those with higher IQs are better chess players when they begin, but after a few years of practice, the relationship is much weaker. One study estimated that only around 12 per cent of the variation in chess player rankings could be attributed to IQ (Grabner et al, 2007). Indeed, measures of general ability account for only around 4 per cent of the variation in the performance of experts (Ericsson et al, 2006). David Berliner (1994) has shown that expertise in teaching appears to be very similar to expertise in other areas so, as noted above, a strategy of getting 'the best and the brightest' into teaching is not only not sufficient to build an outstanding teacher workforce, it is not even necessary.

What does produce expertise is at least ten years of deliberate practice – an effortful focus on improving performance (Ericsson, 2002). Most studies of the effects of experience on teachers' productivity find that teachers improve for the first two or three years, but most slow down, and many stop improving after this (Rivkin, Hanushek and Kain, 2005). This suggests that many teachers are only scratching the surface of the kinds of improvements that are possible.

If we are to help teachers gain the expertise that the research suggests is possible, then first we need to recruit those with a passion for the job. Deliberate practice is not inherently enjoyable – it is instrumental in achieving further increases in performance – and only those who are passionate about helping all students achieve at high levels will be willing to invest the energy needed.

Then, we need to create environments in which all teachers embrace the idea of continuous improvement. This is not the hackneyed idea of 'keeping up with new developments' – it is, rather, an acceptance that the impact of education on the lives of young people creates a moral imperative for even the best teachers to continue to improve. The evidence from studies of focused attempts to improve the performance of serving teachers (Wiliam et al, 2004; Allen et al, 2011) is that the effects of such can be two

or three times as great as the combined effect of all the attempts to improve teaching by teacher replacement outlined above.

Once it is accepted that the benefits of education, for both individuals and for society, create a moral imperative for all teachers to improve, the next step is to decide how this is to be achieved. Obviously teachers could be left to their own devices to improve but, given the importance of educational achievement, it seems it would be important to ensure that teachers should be supported in improving their practice, and the most obvious way to do this is through the provision of feedback.

Section 7: Feedback is more complex than generally assumed

Feedback today has a significant role in evaluating teaching, but it did not originally emerge from the education field. In this section I outline how the idea developed over time, the importance of meta-analysis historically in developing a theory of feedback intervention, and the crucial relationship between type of feedback, context, purpose and use.

The term 'feedback' has its origins in system engineering (Wiener, 1948), and was defined as the 'control of a machine on the basis of its actual performance rather than its expected performance' (Wiener, 1950/1989 p 24). Wiener and his colleagues identified two types of feedback loop – positive and negative – but these terms were used in a technical sense that does not relate in a straightforward way to the way in which we use these terms now.

One example of a positive feedback loop is when a microphone picks up the sound from a loudspeaker that is then further amplified, which is in turn picked up by the microphone, creating the familiar howl of acoustic feedback. Another example is when, in times of shortage, people hoard scarce supplies, which makes supplies even scarcer, which makes people hoard even more, and so on. The point is that, in engineering, positive feedback is unhelpful, leading to explosive growth or collapse. An example of a negative feedback loop is the thermostat that monitors the temperature in a room and then, if the room temperature deviates too far from the desired temperature, the thermostat activates a heating or cooling system that restores the room to the desired temperature. In engineering then, only negative feedback is helpful, restoring the system to its desired state.

When the idea of feedback was taken up in psychology in the 1960s and 1970s, the field of psychology was dominated by behaviourism, and therefore it is not surprising that it was assumed that the most appropriate use of feedback should be used to reinforce desired behavior.

> *the machine like any private tutor, reinforces the student for every correct response, using the immediate feedback not only to shape behavior most effectively but to maintain it in strength in a manner which the layman would describe as 'holding the student's interest'.*
>
> (Skinner, 1968 p 39)

Responding to this, Kulhavy (1977) suggested the following.

> *With such confident statements available, it is no surprise that scholars have worked overtime to fit the round peg of feedback into the square hole of reinforcement. Unfortunately, this stoic faith in feedback-as-reinforcement has all too often led researchers to overlook or disregard alternate explanations for their data. One does not have to look far for articles that devote themselves to explaining why their data failed to meet operant expectations rather than to trying to make sense out of what they found.*
>
> (p 213)

During the following years, a number of reviews of research on the effects of feedback appeared providing further evidence that feedback that merely provided reinforcement was not particularly effective (see, for example, Bangert-Drowns et al, 1991).

Many of the reviews of research on the effects of feedback used meta-analysis to synthesise the results of different studies. Meta-analysis is a technique that expresses the strength of a finding in the form of a standard measure, such as the standardised effect size (Cohen, 1988), and can be a very useful way of synthesising results from different studies. However, a number of cautions need to be borne in mind when looking at the results of meta-analyses, particularly in the field of educational research.

The file drawer problem

There is a well-documented bias in favour of the publication of studies that find significant rather than non-significant results. This is, of course, understandable, but what is less widely realised is that most research studies in education and psychology have relatively low statistical power, generally because the experiments are too small to generate statistically significant results consistently, even if the phenomena they are investigating are real. One investigation estimated that the average statistical power of the typical psychology experiment was around 0.4, meaning that a given experiment had only a 40 per cent chance of producing a statistically significant result, even if the effect under study was real. Because of this, only studies that, by chance, show a larger effect than average are likely to be published, so that the effect sizes of studies that are published are an over-estimate of the true-effect.

Variation in variability

While psychological studies often look at relatively stable phenomena, such as personality, educational studies are generally more interested in change. This is a particular problem for meta-analysis since the effect of an intervention will depend on the variability of the population. The most common measure of effect size, the standardised effect size mentioned above, is calculated by dividing the difference in the mean of the treatment and the control group in an experiment by the standard deviation of the population under study. So if the population under study is a sub-set of some larger population (students with special needs, or gifted students for example), then the denominator of the fraction in the effect size calculation is reduced, so the resulting estimate of the effect size is increased. This is a particular problem when we look at students of different ages. Bloom et al (2008) found that one year's growth for a 6-year-old was equivalent to 1.5 standard deviations, while for a 15-year-old, one year's growth was only 0.2 standard deviations. An experiment involving an intervention that increased the rate of learning by 50 per cent would therefore be expected to have an effect size of 0.75 if it was conducted on six-year-olds, but an effect size of only 0.1 if it was conducted on 15-year-olds.

Selection of studies

The selection of studies for inclusion in a meta-analysis involves a number of decisions, of varying degrees of subjectivity. Ruiz-Primo and Li (2013) reviewed over 9000 papers potentially of relevance to the effectiveness of feedback in the learning of mathematics, science and technology. In the 238 papers they retained, 95 papers had specific quantitative findings on the effects of feedback on the learning of mathematics and science. However, of these 95 papers, 76 of them involved a single feedback event lasting a matter of minutes. While such findings may be of interest to researchers, it is highly unlikely that such results generalise in a straightforward manner to the ongoing effects of feedback over weeks and months.

As noted above, this does not mean that meta-analysis is useless as a technique for aggregating research findings from multiple sources, but it does make clear that considerable caution is needed in drawing conclusions from effect sizes, particularly in terms of the magnitude of effects of different interventions.

A particularly important meta-analysis of feedback studies was conducted by Kluger and DeNisi (1996) in which they reviewed all the studies that had been conducted on the effects of feedback from 1905 to 1995.

They began by defining feedback interventions (FI) as 'actions taken by (an) external agent(s) to provide information regarding some aspect(s) of one's task performance' (p 255). They identified over 3000 relevant studies published between 1905 and 1995, but noted that many of these were very small studies (in many cases involving only a single participant), and were reported in insufficient detail to permit the calculation of an effect size for the intervention. In order to be sure that poor-quality studies were not being included, Kluger and DeNisi established three criteria for inclusion in their review.

1. The participants had to be divided into two groups, the only difference between the groups, as far as could be judged, being whether they received feedback or not.
2. The study involved at least ten participants.
3. They included a measurement of performance with sufficient details provided for the size of the impact of feedback on performance to be calculated.

Surprisingly, only 131 of the 3000 relevant studies satisfied these criteria. These selected studies reported 607 effect sizes, involving 23,663 observations of 12,652 participants. Across all the studies, the average effect size for feedback was 0.41 standard deviations, but the effects varied considerably across the different studies. Most notably, 38 per cent of the 607 effect sizes were negative. In other words, in almost two out of every five cases, feedback actually lowered average performance. In seeking to understand this, they looked for 'moderators' of feedback effects (variables that can explain the differences in effects in different studies) and found that feedback interventions were least effective when they focused attention on the self, more effective when they focused on the focal task, and most effective when they focused on the details of the focal task and when they involved goal-setting.

However, they concluded that whether feedback 'works' or not, and if so, by how much, were not the right questions to ask.

> *Before we conclude, we must reflect on the applied implication of our study. The identification of a number of moderators suggests that in certain situations, FI [feedback intervention] can yield a large and positive effect on performance. Specifically, an FI provided for a familiar task, containing cues that support learning, attracting attention to feedback-standard discrepancies at the task level (velocity FI and goal setting), and is void of cues to the meta-task level (eg, cues that direct attention to the self) is likely to yield impressive gains in performance, possibly exceeding 1 SD. However, even such an FI is not always an efficient intervention. Even when FI has considerable positive effects on performance, its utility may be marginal or even negative. When an FI increases performance through an increase in task motivation, the effect may depend on a continuous FI. Removal of such an FI may result in a reversal as some field studies have demonstrated (Komaki et al, 1980). Therefore, the cost of maintaining a continuous intervention should be considered in evaluating such an intervention. If, however, FI affects performance through task-learning processes, the effect may create only shallow learning and interfere with more elaborate learning. Lack of elaborate learning affects the ability to use the learned material in transfer tasks where the task is similar but not identical (eg, Carroll and Kay, 1988). Moreover, the evidence for any learning effect here was minimal at best. Indeed, in the MCPL [multiple-cue probability learning paradigm] literature, several reviewers doubt whether FIs have any learning value (Balzer et al, 1989; Brehmer, 1980) and suggest alternatives to*

> *FI for increasing learning, such as providing the learner with more task information (Balzer et al, 1989). Another alternative to an FI is designing work or learning environments that encourage trial and error, thus maximizing learning from task feedback without a direct intervention (Frese and Zapf, 1994). These considerations of utility and alternative interventions suggest that even an FI with demonstrated positive effects on performance should not be administered whenever possible. Rather, additional development of FIT [feedback intervention theory] is needed to establish the circumstance under which positive FI effects on performance are also lasting and efficient and when these effects are transient and have questionable utility. This research must focus on the processes induced by FIs and not on the general question of whether FIs improve performance – look at how little progress 90 years of attempts to answer the latter question have yielded.*
>
> (p 278)

In other words, any attempt to understand the effects of feedback without considering how the recipient reacts to the feedback is doomed to fail, because feedback given to one individual might be effective, but the same feedback might be ineffective for a very similar individual, because of the way in which the individuals concerned react to the feedback.

To address this issue, Kluger and DeNisi proposed a 'preliminary feedback intervention theory' based on the observation that there are two situations in which feedback can be provided (those where current performance falls below desired performance, and those where current performance exceeds desired performance) and there are four responses that an individual can make to the feedback (change behaviour, change the goal, abandon the goal, or reject the feedback). This leads to the eight possible effects of feedback interventions shown in Table 1.

In other words, there are eight possible responses to a feedback intervention, and six of them are likely to be ineffective or worse. Only two responses, highlighted in bold in Table 1, are likely to have positive outcomes. Crucially, the effects of feedback depend on the context in which it is given. Therefore, in order to understand how feedback can improve teacher performance, feedback needs to be embedded in a wider theoretical framework that includes both the role of those who are giving feedback and those who are receiving feedback.

There are many ways in which this could be achieved, and discussion of even a few of these is well beyond the scope of this paper. In what follows, one model for the improvement of learning is explored in detail in the context of the formative evaluation of teacher performance.

As an OECD report on teacher evaluation (Santiago and Benavides, 2009) notes, teacher evaluation typically has two main purposes: to improve teacher performance and to provide evidence with which to hold teachers and educational institutions (eg, schools, districts, states) to account. The report also notes that there is a fundamental tension between these two purposes. For example where ratings of

Table 1. Responses to feedback interventions (Kluger and DeNisi, 1996)

Response type	Feedback indicates performance exceeds goal	Feedback indicates performance falls short of goal
Change behaviour	Exert less effort	**Increase effort**
Change goal	**Increase aspiration**	Reduce aspiration
Abandon goal	Decide goal is too easy	Decide goal is too hard
Reject feedback	Feedback is ignored	Feedback is ignored

teaching performance are linked to decisions about job tenure or financial rewards, teachers are unlikely to try innovative approaches, and will organise their lessons to minimise the chances that weak areas of practice are revealed. What is less obvious is that as well as being in tension in terms of how they operate, the different functions of assessment of teacher performance need to be validated in different ways. In the next section, the nature of formative assessment is explored in further detail, while in the following section I discuss the validity of formative evaluations of teacher performance. The paper concludes with some recommendations for ways of implementing formative evaluation of teacher performance.

Section 8: Formative evaluation of teacher performance

There is no widely agreed definition of what, exactly, constitutes formative evaluation. The distinction between formative and summative evaluation was first made by Michael Scriven, in the context of curriculum evaluation. On the one hand, he pointed out that evaluation 'may have a role in the on-going improvement of the curriculum' (Scriven, 1967, p 41) while, in another role, evaluation 'may serve to enable administrators to decide whether the entire finished curriculum, refined by use of the evaluation process in its first role, represents a sufficiently significant advance on the available alternatives to justify the expense of adoption by a school system' (p 41–42). He then proposed 'to use the terms 'formative' and 'summative' evaluation to qualify evaluation in these roles' (p 43). The same distinction was then applied by Benjamin Bloom to the evaluation of individual students:

> *Quite in contrast is the use of 'formative evaluation' to provide feedback and correctives at each stage in the teaching–learning process. By formative evaluation we mean evaluation by brief tests used by teachers and students as aids in the learning process. While such tests may be graded and used as part of the judging and classificatory function of evaluation, we see much more effective use of formative evaluation if it is separated from the grading process and used primarily as an aid to teaching.*
>
> (Bloom, 1969, p 48)

Since then, a number of authors have provided a range of definitions of the term 'formative assessment' (see Wiliam, 2011b, for an extended discussion). The principal sources of variation amongst these definitions are

- the duration of the interval between the collection of the evidence of achievement and its use;
- whether it is essential that the students from whom evidence was elicited are beneficiaries of the process;
- whether the assessment has to change the intended instructional activities; and
- whether students have to be actively engaged in the process.

we see much more effective use of formative evaluation if it is separated from the grading process and used primarily as an aid to teaching

In attempting to provide a comprehensive definition of formative assessment, Black and Wiliam (2009) proposed an inclusive definition that encompassed all of the various issues identified above as variations of a central idea. Paraphrased for the formative evaluation of teacher performance, Black and Wiliam's definition is as follows.

> *An evaluation of teacher performance functions formatively to the extent that evidence of teacher performance elicited by the assessment is interpreted by leaders, teachers, or their peers to make decisions about the professional development of the teacher that are likely to be better, or better founded, than those that would have been taken in the absence of that evidence.*

Several features of this definition merit explanation. The first is that the definition of formative assessment rests on the function that the assessment serves rather than on the nature of the assessment itself. Since any assessment can function formatively or summatively, there can be no such thing as 'a formative assessment' but, rather, only an assessment that elicits evidence that is used formatively. Second, the focus of the definition is on decisions rather than the intentions behind the data collection. This is to ensure that situations in which evidence is collected with the intention of improving teacher performance, but where the evidence is not actually used, would not be regarded as formative. In other words, the focus is on decision-driven data collection rather than data-driven decision making. Third, the definition is silent about who (ie, leaders, the teachers themselves, or their peers) make the decisions (the term 'leader' is used here for anyone who is professionally responsible for the professional development of teachers, whether they are called leaders, coaches, mentors, and so on). Fourth, the definition does not require that the process actually improves the teacher's professional development – given the complexity of human learning, there can be no such guarantees. However the definition does require that the resulting decisions are likely to improve that teacher's learning or, and this is the fifth point, the definition allows for situations in which the evidence confirms that the actions that would have been taken in the absence of the evidence were, in fact the correct ones. In such a situation, the decisions taken are not better decisions (because they are precisely the same decisions) but they are better founded, being based on firmer evidence.

From the definition, it can be seen that formative evaluation is concerned with the creation of, and capitalisation upon, 'moments of contingency' in teachers' learning for the purpose of the regulation of the teacher's learning processes. This might seem to be a very narrow focus, but it helps to distinguish a theory of formative assessment from an overall theory of teaching and learning. However, whilst this focus is narrow, its impact is broad, since how teachers, learners, and their peers create and capitalise on these moments of contingency entails considerations of instructional design, curriculum, pedagogy, psychology and epistemology. In the following section, the central idea of formative evaluation is analysed in more detail and particularly in terms of related strategies.

Section 9: Strategies of formative evaluation of teacher performance

Since formative evaluation of teacher performance is essentially concerned with the regulation of teachers' learning processes, one way to think about this is that it is concerned with three central processes. These are

1. the goal for the teacher's learning;
2. their current level of performance; and
3. the steps needed to reach the goal.

If we consider these three processes in conjunction with the roles of the various individuals involved in this process – the teacher, the teacher's peers, and those who are professionally responsible for that teacher's learning, for convenience here termed 'leaders' – crossing these two dimensions leads to a 3x3 matrix of cells. The contents of each of the nine cells could be discussed individually, but the model is considerably simplified if we group some of the cells together, as shown in Table 2, which is modified from Wiliam and Thompson (2008). Each of the five strategies is discussed below in turn.

Clarifying, understanding, and sharing learning intentions

Perhaps the most problematic aspect of formative evaluation of teaching performance relates to points made earlier, in Section 5. Because we have little idea about the

Table 2. Aspects of formative assessment

<table>
<tr><th></th><th>Where the teacher is going</th><th>Where the teacher is right now</th><th>How to get there</th></tr>
<tr><td>Leader</td><td rowspan="3">Clarifying, understanding and sharing learning intentions and criteria for success</td><td>Engineering effective situations, activities and tasks that elicit evidence of development</td><td>Providing feedback that moves teachers forward</td></tr>
<tr><td>Peer</td><td colspan="2">Activating teachers as learning resources for one another</td></tr>
<tr><td>Teacher</td><td colspan="2">Activating teachers as the owners of their own learning</td></tr>
</table>

(After Wiliam and Thompson, 2008)

characteristics of effective teaching practice, it is difficult to make sure that the formative evaluation of teachers is appropriately directed. Put simply, if we do not know what good teaching looks like, how can we improve teachers? More importantly, given that our observation frameworks capture only a small proportion of the variation in teacher quality, there is a real danger of establishing goals for teacher development that actually make teachers less effective.

In recent years, considerable attention has been focused on the use of rubrics to communicate standards to learners, both for school students and for teaching performance. All the main teacher evaluation models such as Danielson's 'Framework for Teaching', mentioned above, and the teacher evaluation model developed by Marzano and Toth (2013) present levels of teacher performance in the form of scoring rubrics that identify different levels of performance in the area in question. Presenting levels of performance in the form of a scoring rubric can undoubtedly be useful, but it is important to note that scoring rubrics may not be effective in communicating levels of performance to those that are not already able to demonstrate them. Rubrics are often treated as if they were instructions about how to improve performance but they are more often post hoc descriptions of quality. More importantly, while experts are often able to identify what they are doing with the descriptions contained in rubrics, the contents of the rubrics are not used by experts in their performance. As Michael Polanyi wrote

> *Maxims are rules, the correct application of which is part of the art which they govern. The true maxims of golfing or of poetry increase our insight into golfing or poetry and may even give valuable guidance to golfers and poets; but these maxims would instantly condemn themselves to absurdity if they tried to replace the golfer's skill or the poets art.* ***Maxims cannot be understood, still less applied by anyone not already possessing a good practical knowledge of the art. They derive their interest from our appreciation of the art and cannot themselves either replace or establish that appreciation.***
>
> (Polanyi, 1958 p 31–32, my emphasis)

Rubrics may therefore provide a valuable starting point for conversations between teachers and their leaders, but slavish adherence to the text of the rubrics is unlikely to improve teaching. Because of the importance of context, it is likely that examples of actual practice, ideally on video, together with commentary that draws out significant features, is likely to be far more effective in communicating to teachers aspects of high-quality performance.

Engineering effective situations, activities and tasks that elicit evidence of development

The research on the generalisability of ratings of teacher performance discussed in Section 5 suggests that any one observation of teaching performance is unlikely to yield robust evidence of a teacher's capabilities. This is obviously a significant problem for the accountability function of assessment, since the performance observed on any one occasion is not a reliable indicator of the teacher's performance on another occasion. However, for the improvement function of assessment, the variability of teacher performance can be useful, since observations of teaching can be scheduled for specific occasions when the observation of teaching practice is most likely to be beneficial for the teacher's development. In general, therefore, this suggests that it should be the teacher being observed who should choose the lesson to be observed. One important point to bear in mind is that all observations are theory-dependent. Even in physics, as Werner Heisenberg noted, 'What we learn about is not nature itself, but nature exposed to our methods of questioning.' (Johnson, 1996, p 147) For the observer of the teaching practice, it is likely to be helpful to have an opportunity to meet with the teacher before the lesson, in order to understand what the teacher is seeking to do, and to have a significant period of time after the lesson to try to understand the teacher's own understanding of what happened in the lesson. As David Ausubel remarked many years ago,

> *If I had to reduce all of educational psychology to just one principle, I would say this: 'The most important single factor influencing learning is what the learner already knows. Ascertain this and teach him accordingly.'*
>
> (Ausubel 1968, p vi)

the quality of the relationship between those giving and receiving feedback is crucial in determining whether feedback has a positive effect.

This would appear to be as true for the learning of teachers as it is for the learning of school students.

The idea that the teacher being observed should choose the lesson to be observed was a particularly significant feature of the My Teaching Partner coaching system (Allen et al, 2011). This focused attention on three aspects of teaching, which were:

1. emotional support for students (positive relationships, teacher sensitivity and regard for adolescent perspectives);
2. classroom organisation (behaviour management, maximising learning time and effective instructional formats); and
3. instructional support (content understanding, analysis and problem solving, and quality of feedback).

Every two weeks, participating teachers video-recorded one lesson, and uploaded the recording to a secure server where the coach could review the video and select a small number of short segments (one to two minutes in duration) for detailed discussion via telephone. After two years, students taught by teachers participating in the My Teaching Partner system were learning 50 per cent more than those taught by matched teachers not participating in the program.

Providing feedback that moves teachers forward

As is clear from the extended discussion of feedback in Section 7 above, the quality of the relationship between those giving and receiving feedback is crucial in determining whether feedback has a positive effect. Leaders need to know their teachers so they know when to be critical and when to provide support. Just as importantly, teachers need to trust their leaders, because unless the teacher believes that the leader has the teacher's best interests at heart, and unless the teacher believes the leader has credibility as a coach, the teacher is unlikely to invest the effort needed to improve practice. This means that there can be no

simple recipe for effective feedback for teachers on their teaching performance, but a couple of principles derived from other research on feedback may be useful here. The first is that feedback should cause thinking. Feedback that causes an emotional reaction, as is likely when the feedback compares an individual teacher's performance with that of other teachers, is unlikely to be helpful. Far more helpful is likely to be comparing a teacher's performance with her/his own previous performance (in other words, was this a 'personal best' for the teacher?) which is likely to help the teacher adopt a 'growth mindset' (Dweck, 2006). The second principle is that feedback should be more work for the recipient than the donor. The feedback event itself is likely to be relatively unimportant in improving teacher performance; what matters is the follow-up action taken by the teacher.

Activating teachers as learning resources for one another

Because, as noted above, the issue of trust between the donor and the recipient of feedback is crucial to the likely success of the feedback (see also Santiago and Benavides, 2009), it may be helpful to involve peers, rather than those with a formal leadership role within the school or district, in providing feedback to teachers. This is particularly important where leaders have a formal role in the accountability function because it can be difficult for leaders to separate out the two roles, and, for example, ignore evidence that might be relevant to the accountability function if they are meant to be focusing on the improvement function. Even if leaders are able to do this, ultimately the behaviour of the teacher will depend not on whether the leader is able to separate these two roles clearly, but whether the teacher believes the leader is able to do so. If the teacher believes that evidence of weaknesses in practice, revealed in an observation that is ostensibly intended to improve practice, may affect the judgement made about the teacher's effectiveness, then the teacher is more likely to 'play safe' – so that the potential for the observation to improve practice is reduced. Where peers are involved in classroom observation, it can be particularly helpful to have a clear protocol for the lesson observation, which makes clear that

- the teacher being observed specifies the focus of the observation;
- the teacher being observed specifies the evidence to be collected; and
- the teacher being observed owns any notes made by the observer during the lesson.

By making it clear that the teacher being observed 'owns' the process, such an observation is clearly distinguished from observations for accountability purposes and thus makes it easier for a trusting relationship to be developed.

Activating teachers as the owners of their own learning

Ultimately, the amount of time for leaders and peers to observe practice will be limited so, if improvement is to occur, most of it will be generated by the teacher's own efforts to improve. Some have argued that this is best accomplished through systems of incentives – particularly financial incentives – for teachers, but the evidence, both in the teaching profession and more widely, suggests that performance-related pay is not particularly successful in improving performance (Pfeffer, 1998; Springer et al, 2010). A more likely route for teacher improvement comes from engagement with the moral imperative identified in Section 1, and the realisation that, as shown in Section 2, teachers can make a difference. While estimates of the relative magnitudes of different influences on student learning are fraught with difficulties, and may vary considerably from culture to culture and country to country, there is now substantial evidence that the impact of teacher effects at least rival, and may well exceed, the impact of family background and socio-economic status (Rowe, 2003). When teachers do their job better, their students live longer, are healthier, and contribute more to society.

Currently, in many systems, professional support is seen as something that is needed only by the weakest practitioners, so that being offered professional support can be perceived as an indication of poor performance. When, instead, all teachers embrace the idea that they can improve, not because they are not good enough, but because they can be even better, this creates a natural collegiality that supports all teachers in embracing the need for continuous improvement. As the research reviewed in Section 6 shows, most teachers slow, and many teachers stop improving after two or three years in the job, so the expertise research suggests that considerable improvements are possible if all teachers, rather than the weakest, engage in continuous professional improvement.

Section 10: The validity of formative evaluation of teacher performance

Once systems adopt the formative evaluation of teacher performance, an immediate concern is with the quality of such evaluation. Traditionally, in evaluation, concerns about quality have been addressed through the concept of validity. Originally conceptualised as a property of a test (ie, a test is valid to a certain degree, or not), it has become accepted that validity only makes sense as a property of inferences based on assessment outcomes.

so the expertise research suggests that considerable improvements are possible if all teachers, rather than the weakest, engage in continuous professional improvement

For example, a mathematics test with a high reading demand might support conclusions about the mathematical abilities of good readers but, for poor readers, we would not know whether poor performance was due to weak mathematical capability, or poor reading. The test would support some inferences (eg, about the mathematical capability of good readers) but not others (eg, about the mathematical capability of poor readers). As Lee Cronbach noted, 'One validates, not a test, but **an interpretation of data arising from a specified procedure'** (Cronbach, 1971, p 447, emphasis in original).

For the accountability function of assessment, the inferences we wish to draw are generally about the quality of teaching observed, and one aspect that is particularly important is that such inferences are free from subjectivity – in other words we want to be assured that, notwithstanding the poor generalisability of teacher ratings, noted above, the rating given to a teacher does not depend too greatly on the person who made the rating. Given the complexity of practice, it is unlikely that such a rating could ever be truly criterion-referenced (ie, requiring only the application of specified criteria) but such judgements can, through training of observers, be free from subjectivity. Where different raters agree on the quality of what they have observed, we might say that the assessment is construct-referenced (Wiliam, 1994) relying on the shared construct of quality in the minds of those making the judgements. In other words, the meanings of assessment outcomes should be common across different assessors. On the other hand, where improvement is the main goal, consistency of meanings across interpreters is much less important. If two different assessors interpret a particular teaching performance differently, and suggest different professional development activities that would be equally successful in moving the teacher's learning forward, then, according to the definition of formative evaluation adopted in Section 8, they would be equally valid in terms of their impact on teacher learning. In other words, adopting a distinction used by Samuel Messick (1988), if summative functions of assessment are validated by their meanings, then formative functions of assessment are validated by their consequences (Wiliam and Black, 1996).

Before leaving the issue of the validity of formative assessment, one further comment is warranted. As noted above, for assessment to function formatively, teachers need to become members of the same community of practice of which their leaders are already members – they need to share the implicit concepts of quality that raters share when they agree about quality. As Royce Sadler stated, in the context of students and teachers,

> *The indispensable conditions for improvement are that the student comes to hold a concept of quality roughly similar to that held by the teacher, is able to monitor continuously the quality of what is being produced during the act of production itself, and has a repertoire of alternative moves or strategies from which to draw at any given point. In other words, students have to be able to judge the quality of what they are producing and be able to regulate what they are doing during the doing of it.*
>
> (p 121)

While it may be something of an overstatement to say that these conditions are indispensable, it seems likely that the same ideas would be very powerful indicators of effective teacher learning.

Section 11: Implementing formative evaluation of teacher performance

A commitment to the formative evaluation of teacher performance does not provide any indication about how this should be done. There are many different models that might be adopted, and each will have strengths and weaknesses relative to the institutions in which they are to be used. However, as a result of extensive work over a 15-year period with teachers developing classroom practice, five principles of teacher learning appear to be particularly important (Wiliam, 2012). They are choice; flexibility; small steps; accountability; and support. Each of these is discussed in turn below.

Choice

When reporting to teachers on the results of observations of their practice, it is common to report back in terms of 'strengths' and 'areas for development.' The use of the term 'areas for development' is presumably meant to make criticism more palatable but the effect is to create an unfortunate implication that weaknesses are necessarily areas for development. Of course for some teachers, weaknesses may indeed be areas for development, but for other teachers they may just be weaknesses. The important point for formative evaluation of teacher performance is not the profile of strengths and weaknesses but which areas of a teacher's performance, if developed, would have the biggest impact on student learning.

In the business world, there has been a growing realisation over the last thirty or so years that organisations can often benefit more by having individuals become truly outstanding at the things they are already good at rather than worrying too much about weaknesses (Belbin, 1981; Buckingham, 2007). In the same way, the aim of professional development for teachers should not be to make every teacher into a clone of every other teacher, but to help each teacher become the best teacher s/he can be. For some teachers, this may well require a focus on weaknesses but for most, it seems likely that a focus on strengths will be more successful.

Flexibility

As well as having some say over what they will develop, it is also important that teachers have flexibility in 'morphing' (Ginsburg, 2001) ideas that they encounter, to make them work in their own classrooms. The problem is that when teachers adapt ideas derived from research findings to make them work in their own classrooms, they often distort them so greatly that they are no longer effective in improving student achievement. If we are to give teachers freedom to adapt the ideas that they encounter, to make them work in their own classrooms, we also have to provide strong frameworks to ensure that the changes they make do not

render the changes ineffective. For example, it is well established (see, eg, Slavin, 1995) that collaborative learning is an effective way of raising student achievement, provided that the way the instruction is designed requires group goals (so that students are working as a group rather than just in a group) and individual accountability (so that the failure of an individual impacts upon the whole group). A survey of 85 elementary school teachers found that 93 per cent of the teachers said they employed collaborative learning. However, follow-up interviews with 21 of the teachers revealed that just five teachers implemented collaborative learning in such a way as to create both group goals and individual accountability (Antil et al, 1998).

Small steps

Given the moral imperative for improving teaching, described in Sections 1 and 6, it is not surprising that policymakers, politicians and administrators want to get teachers developing their classroom practice as quickly as possible. However, research on the impact of professional development suggests that the benefits have been disappointing – what Michael Fullan said over twenty years ago seems as true today as it was then.

> *Nothing has promised so much and has been so frustratingly wasteful as the thousands of workshops and conferences that led to no significant change in practice when teachers returned to their classrooms.*
>
> (Fullan and Stiegelbauer, 1991, p 315)

Some have attributed the slowness of teacher change to resistance on the part of teachers – suggesting teachers cling to a set of professional habits that represent a core part of each teacher's professional identity, which is why they are unwilling to change. Such beliefs provide the rationale for incentive schemes – the idea is that teachers will adopt new ideas if they are paid to do so. However, as we saw in Section 6, financial incentives do not appear to have been effective in improving teacher performance.

A more likely explanation of the slowness of teacher change comes from the research on expertise discussed in Section 6. Expertise is the result of extensive deliberate practice, and the development of expertise cannot be short-circuited by telling teachers 'what to do'. Professional development amounts to the acquisition of new aspects of expertise, which takes time.

Accountability

As was shown in Section 2, educational outcomes depend on a range of factors, many of which are beyond the control of schools and teachers. Holding schools and teachers accountable for things that they cannot influence would seem to be contrary to principles of natural justice. However, what every teacher does control is whether s/he improves or not. That is why the improvement function of assessment is more important than the accountability function. We can spend a lot of time and energy seeking to measure the quality of teachers, but even if we could do this well, and even if we could then remove the least effective, the benefits would be modest. A focus on accountability for improvement will have far greater impact (because all teachers would then be improving).

Of course there are many different protocols that might be adopted for action planning but, in my work with teachers over the last 15 years, I have found it helpful to engage them in a highly structured planning approach that emphasises four processes.

1. The plan should identify a small number of changes that the teacher will make in her/his teaching: When teachers try to change more than two or three things in their practice at the same time, the result is often that their classroom routines deteriorate significantly, and they then fall back on those routines with which they feel comfortable or 'safe'.
2. The plan should be written down: Writing the plan down makes it more likely that the teacher thinks the plan through while writing it down. It makes the ideas more

concrete and also creates a record that means teachers are less likely to forget what they planned to do.

3. The plan should focus on areas of practice likely to benefit students: Not all changes that teachers make to their practice will benefit students. Given the moral imperative for the improvement of education, teachers should focus on aspects of their practice that are likely to improve outcomes for their students. Because research evidence rarely provides evidence that is applicable to all contexts, teachers will need to use their professional judgement in deciding where to focus their efforts, but they should be able to provide some evidence that what they choose to work on does have at least a prima facie case for impact on outcomes for their students.
4. The action plan should identify what the teacher plans to reduce, or give up doing to make time for the changes: most teachers are working as hard as they can, so if these changes are treated as an addition to their load, they are unlikely ever to be implemented. To make time for these changes, the action plan must identify something that the teachers are currently doing that they will stop doing, or do less of, to make time available for the changes. Asked to make such clear priorities, people often hope that they can make the necessary changes by being more efficient in their use of time, but this is usually hopelessly optimistic. The only way to make time available for new things is to reduce, or stop doing entirely, things that they are currently doing, in order to create time for innovation. Moreover, in education, almost everything that teachers do benefits students. If teachers look for unproductive activities that they can discontinue in order to make time for improvement, they will not find any. In a very real sense, the essence of leadership in education is stopping people doing good things to give them time to do even better things.

Support

The last process element – support – is closely related to accountability. Indeed, some authors have described them as a single feature of effective learning environments for teachers: supportive accountability (Ciofalo and Leahy, 2006). The central idea is the creation of structures that, while making teachers accountable for developing their practice, also provide the support for them to do this. Support and accountability can therefore be thought of as two sides of the same coin. As noted above, the teachers' role is to commit to improvements in their practice, and to focus on changes that are likely to benefit their students. Leaders can create effective learning environments for their teachers by creating expectations for continually improving practice, by keeping the focus on the things that make a difference to students, by providing the time, space, dispensation and support for innovation and, finally, by supporting teachers in taking risks.

Conclusion

In this paper I have reviewed research on various strategies for the improvement of educational outcomes for school students, and concluded that investment in teachers who are currently serving must be the central strategy of any serious attempt to improve schooling. To be of maximum effect, investment in teachers must build on the evidence about what kinds of changes in teacher practice have the biggest impact on student achievement, but it must also take into account how teachers learn and develop. The evidence presented in this paper suggests that substantial improvements in educational outcomes for young people, with the attendant benefits for individuals and society, are possible if we focus on the power of assessment to improve, rather than measure, teacher performance.

References

Allen, J P, Pianta, R C, Gregory, A, Mikami, A Y and Lun, J (2011) 'An interaction-based approach to enhancing secondary school instruction and student achievement', *Science*, **333**, 6045, p 1034–1037.

Antil, L R, Jenkins, J R, Wayne, S K and Vadasy, P F (1998) 'Cooperative learning: Prevalence, conceptualization and the relation between research and practice', *American Educational Research Journal*, **35**, 3, p 419–454.

Atteberry, A, Loeb, S and Wyckoff, J (2013) *Do First Impressions Matter? Improvement in Early Career Teacher Effectiveness*, Center for Analysis of Longitudinal Data in Educational Research, Washington, DC.

Ausubel, D P (1968) *Educational Psychology: A Cognitive View*, Holt, Rinehart and Winston, New York.

Bangert-Drowns, R L, Kulik, C-L C, Kulik, J A and Morgan, M (1991) 'The instructional effect of feedback in test-like events', *Review of Educational Research*, **61**, 2, p 213–238.

Barber, M and Mourshed, M (2007) *How the World's Best-performing School Systems Come Out on Top*, McKinsey and Company, London.

Belbin, R M (1981) *Management Teams: Why They Succeed or Fail*, Heinemann, Oxford, UK.

Berliner, D C (1994) 'Expertise: The wonder of exemplary performances', in J N Mangieri and C C Block (Eds) *Creating Powerful Thinking in Teachers and Students: Diverse Perspectives* (p 161–186), Harcourt Brace College, Fort Worth, TX.

Black, P J and Wiliam, D (2009) 'Developing the theory of formative assessment', *Educational Assessment, Evaluation and Accountability*, **21**, 1, p 5–31.

Bloom, B S (1969) 'Some theoretical issues relating to educational evaluation', in R W Tyler (Ed) *Educational Evaluation: New Roles, New Means: The 68th Yearbook of the National Society for the Study of Education (Part II)*, **68**, 2, p 26–50, University of Chicago Press, Chicago.

Bloom, H S, Hill, C J, Black, A R and Lipsey, M W (2008) 'Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions', *Journal of Research on Educational Effectiveness*, **1**, 4, p 289–328.

Bryk, A S, Sebring, P B, Allensworth, E and Luppescu, S (2010) *Organizing Schools for Improvement: Lessons from Chicago*, University of Chicago Press, Chicago.

Buckingham, M (2007) *Now Go Put Your Strengths to Work*, Simon and Schuster, New York.

Ciofalo, J and Leahy, S (2006) 'Personal action plans: Helping to adapt and modify techniques', a paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April.

Cohen, J (1988) *Statistical Power Analysis for the Behavioral Sciences*, Lawrence Erlbaum Associates, Hillsdale, NJ.

Cronbach, L J (1971) 'Test validation', in R L Thorndike (Ed) *Educational Measurement*, 2nd ed, p 443–507, American Council on Education, Washington DC.

Danielson, C (1996) *Enhancing Professional Practice: A Framework for Teaching*, ASCD, Alexandria, VA.

Dobbie, W (2012) *Teacher Characteristics and Student Achievement: Evidence from Teach for America*, Harvard University Faculty of Arts and Sciences, Cambridge, MA.

Dweck, C S (2006) *Mindset: The New Psychology of Success*, Random House, New York.

Ericsson, K A (2002) 'Attaining excellence through deliberate practice: Insights from the study of expert performance', in M Ferrari (Ed) *The Pursuit of Excellence through Education*, p 21–55, Lawrence Erlbaum Associates, Mahwah, NJ.

Ericsson, K A, Charness, N, Feltovich, P J and Hoffman, R R (2006) *The Cambridge Handbook of Expertise and Expert Performance*, Cambridge University Press, Cambridge, UK.

Feinstein, L, Budge, D, Vorhaus, J and Duckworth, K (2008) *The Social and Personal Benefits of Learning: A Summary of Key Research Findings*, Institute of Education, University of London, London.

Fitzpatrick, M D, Grissmer, D and Hastedt, S (2011) 'What a difference a day makes: Estimating daily learning gains during kindergarten and first grade using a natural experiment', *Economics of Education Review*, **30**, 2, p 269–279.

Fullan, M and Stiegelbauer, S (1991) *The New Meaning of Educational Change*, Cassell, London.

Ginsburg, H P (2001) *The Mellon Literacy Project: What Does it Teach Us about Educational Research, Practice, and Sustainability?* Russell Sage Foundation, New York.

Gladwell, M (2008) 'Most likely to succeed', *New Yorker*, 15 December, p 36–42.

Goldhaber, D D, Goldschmidt, P and Tseng, F (2013) 'Teacher value-added at the high-school level: Different models, different answers?' *Educational Evaluation and Policy Analysis*, **35**, 2, p 220–236.

Grabner, R H, Stern, E and Neubauer, A C (2007) 'Individual differences in chess expertise: A psychometric investigation', *Acta Psychologica*, **124**, 3, p 398–420.

Hamre, B K and Pianta, R C (2005) 'Academic and social advantages for at-risk students placed in high quality first grade classrooms', *Child Development*, **76**, 5, p 949–967.

Hanushek, E A (1971) 'Teacher characteristics and gains in student achievement: Estimation using micro data', *American Economic Review*, **61**, 2, p 280–288.

Hanushek, E A (2010) 'Teacher deselection', in D Goldhaber and J Hannaway (Eds) *Creating a New Teaching Profession*, p 165–180, Urban Institute Press, Washington, DC.

Hanushek, E A (2011) 'The economic value of higher teacher quality', *Economics of Education Review*, **30**, 3, p 466–479.

Hanushek, E A and Woessmann, L (2010) *The High Cost of Low Educational Performance: The Long-run Impact of Improving PISA Outcomes*, Organisation for Economic Co-operation and Development, Paris.

Harris, D N and Sass, T R (2007) *Teacher Training, Teacher Quality and Student Achievement*, **3**, National Center for Analysis of Longitudinal Data in Education Research, Washington, DC.

Hill, H C, Charalambous, C Y and Kraft, M A (2012) 'When rater reliability is not enough: Teacher observation systems and a case for the generalizability study', *Educational Researcher*, **41**, 2, p 56–84.

IRPSITEPI (2012) *Report of the International Review Panel on the Structure of Initial Teacher Education Provision in Ireland*, IRPSITEPI report to the Department of Education and Skills, Dublin, Ireland.

Johnson, G (1996) *Fire in the Mind: Science, Faith and the Search for Order*, Viking, London.

Kane, T J and Staiger, D O (2012) *Gathering Feedback for Teaching: Combining High-quality Observations with Student Surveys and Achievement Gains*, Bill and Melinda Gates Foundation, Redmond, WA.

Kluger, A N and DeNisi, A (1996) 'The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory', *Psychological Bulletin*, **119**, 2, p 254–284.

Kulhavy, R W (1977) 'Feedback in written instruction', *Review of Educational Research*, **47**, 2, p 211–232.

Marzano, R J and Toth, M D (2013) *Teacher Evaluation that Makes a Difference: A New Model for Teacher Growth and Student Achievement*, ASCD, Alexandria, VA.

McCaffrey, D, Sass, T R and Lockwood, J R (2008) 'The intertemporal stability of teacher effect estimates', a paper presented at the *National Conference on Value-Added Modeling*, University of Wisconsin, 22–24 April.

McGaw, B (2008) 'The role of the OECD in international comparative studies of achievement', *Assessment in Education: Principles, Policy and Practice*, **15**, 3, p 223–243.

Messick, S (1988) 'The once and future issues of validity: Assessing the meaning and consequences of measurement', in H Wainer and H I Braun (Eds) *Test Validity*, p 33–45, Lawrence Erlbaum Associates, Hillsdale, NJ.

OECD (2013a) *Education at a Glance*, Organisation for Economic Co-operation and Development, Paris.

OECD (2013b) 'What teachers know and how that compares with college graduates around the world', *Education Today*. Accessed 5 September 2014, at oecdeducationtoday.blogspot.com/2013/11/what-teachers-know-and-how-that.html.

Pfeffer, J (1998) 'Seven practices of successful organizations', *California Management Review*, **40**, 2, p 96–124.

Polanyi, M (1958) *Personal knowledge*, Routledge and Kegan Paul, London, UK.

Rivkin, S G, Hanushek, E A and Kain, J F (2005) 'Teachers, schools and academic achievement', *Econometrica*, **73**, 2, p 417–458.

Rockoff, J E and Speroni, C (2011) 'Subjective and objective evaluations of teacher effectiveness: Evidence from New York City', *Labour Economics*, **18**, 5, p 687–696.

Rothstein, J (2010) 'Teacher quality in educational production: Tracking, decay, and student achievement', *Quarterly Journal of Economics*, **125**, 1, p 175–214.

Rowe, K J (2003) 'The importance of teacher quality as a key determinant of students' experiences and outcomes of schooling', a paper presented at the *ACER Research Conference*, Melbourne, October. Accessed 23 November 2013 at www.det.nsw.edu.au/proflearn/docs/pdf/qt_rowe2003.pdf.

Ruiz-Primo, M A and Li, M (2013) 'Examining formative feedback in the classroom context: New research perspectives', in J H McMillan (Ed) *Sage Handbook of Research on Classroom Assessment*, 2nd ed, p 215–232, Sage, Thousand Oaks, CA.

Sadler, D R (1989) 'Formative assessment and the design of instructional systems', *Instructional Science*, **18**, p 119–144.

Santiago, P and Benavides, F (2009) *Teacher Evaluation: A Conceptual Framework and Examples of Country Practices*, Organisation for Economic Co-operation and Development, Paris.

Sartain, L, Stoelinga, S R, Brown, E R, Luppescu, S, Matsko, K K M, Miller, F K, Durwood, C E, Jiang, J Y and Glazer, D (2011) *Rethinking Teacher Evaluation in Chicago: Lessons Learned from Classroom Observations, Principal–teacher Conferences, and District Implementation*, Consortium on Chicago School Research, Chicago.

Scriven, M (1967) 'The methodology of evaluation', in R W Tyler, R M Gagné and M Scriven (Eds) *Perspectives of Curriculum Evaluation*, **1**, p 39–83, Rand McNally, Chicago.

Skinner, B F (1968) *The Technology of Teaching*, Appleton-Century-Crofts, New York.

Slater, H, Davies, N and Burgess, S (2008) *Do Teachers Matter? Measuring the Variation in Teacher Effectiveness in England*, CMPO Working Paper 09/212, University of Bristol Institute of Public Affairs, Bristol. Accessed 5 September 2014, at www.bris.ac.uk/cmpo/publications/papers/2009/wp212.pdf.

Slavin, R E (1995) *Cooperative Learning: Theory, Research and Practice*, 2nd ed, Allyn and Bacon, Boston, MA.

Springer, M G, Ballou, D, Hamilton, L, Le, V-N, Lockwood, J R, McCaffrey, D and Stecher B M (2010) *Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching*, National Center on Performance Incentives at Vanderbilt University, Nashville, TN.

Wiener, N (1948) *Cybernetics, or Control and Communication in the Animal and the Machine*, John Wiley and Sons Inc, New York.

Wiener, N (1950/1989) *The Human Use of Human Beings: Cybernetics and Society*, Free Association Books, London.

Wiliam, D (1994) 'Assessing authentic tasks: Alternatives to mark-schemes', *Nordic Studies in Mathematics Education*, **2**, 1, p 48–68.

Wiliam, D (2011a) 'How do we prepare students for a world we cannot imagine?' a paper presented at the *Salzburg Global Seminar*, Salzburg, Austria, 6 December. Accessed 1 June 2012, at www.dylanwiliam.net.

Wiliam, D (2011b) 'What is assessment for learning?' *Studies in Educational Evaluation*, **37**, 1, p 2–14.

Wiliam, D (2012) *Sustaining Formative Assessment with Teacher Learning Communities*, Kindle Direct Publishing, Seattle, WA.

Wiliam, D and Black, P J (1996) 'Meanings and consequences: A basis for distinguishing formative and summative functions of assessment?' *British Educational Research Journal*, **22**, 5, p 537–548.

Wiliam, D and Thompson, M (2008) 'Integrating assessment with instruction: What will it take to make it work?' in C A Dwyer (Ed) *The Future of Assessment: Shaping Teaching and Learning*, p 53–82, Lawrence Erlbaum Associates, Mahwah, NJ.

Wiliam, D, Lee, C, Harrison, C and Black, P J (2004) 'Teachers developing assessment for learning: Impact on student achievement', *Assessment in Education: Principles Policy and Practice*, **11**, 1, p 49–65.

Winters, M A and Cowen, J M (2013) 'Who would stay, who would be dismissed? An empirical consideration of value-added teacher retention policies', *Educational Researcher*, **42**, 6, p 330–337.

Note

The following references relate to citations embedded in the quotation from Kluger and DeNisi (1996). They were tracked on 7 September 2014, at mario.gsia.cmu.edu/micro_2007/readings/feedback_effects_meta_analysis.pdf.

Balzer, W K, Doherty, M E and O'Connor, R, Jr (1989) 'Effects of cognitive feedback on performance', *Psychological Bulletin*, 106, p 410–433.

Brehmer, B (1980) 'In one word: Not from experience', *Acta Psychologica*, 45, p 223–241.

Carroll, J M and Kay, D S (1988) 'Prompting, feedback and error correction in the design of a scenario machine', *International Journal of Man-Machine Studies*, 28, p 11–27.

Frese, M and Zapf, D (1994) 'Action as the core of work psychology: A German approach', in H C Triandis, M D Dunnette and L M Hough (Eds) *Handbook of Industrial and Organizational Psychology*, 2nd ed, **4**, p 271–340, Consulting Psychologists Press, Palo Alto, CA.

Komaki, J, Heinzmann, A T and Lawson, L (1980) 'Effect of training and feedback: Component analysis of a behavioral safety program', *Journal of Applied Psychology*, **65**, 3, p 261–270.

Additional reading

Although not cited explicitly in the text, the following items were used in preparing this paper, and may be of interest to the reader.

Amabile, T M and Kramer, S J (2011) *The Progress Principle: Using Small Wins to Ignite Joy, Engagement and Creativity at Work*, Harvard Business Review Press, Cambridge, MA.

Goos, M and Manning, A (2007) 'Lousy and lovely jobs: The rising polarization of work in Britain', *Review of Economics and Statistics*, **89**, 1, p 118–133.

Hanushek, E A (2004) *Some Simple Analytics of School Quality*, NBER Working Paper W10229, National Bureau of Economic Research, Washington, DC.

Wiliam, D (2010) 'Standardized testing and school accountability', *Educational Psychologist*, **45**, 2, p 107–122.

CSE/IARTV Publications

Recent titles in the CSE Occasional Papers Series

No. 137 *The formative evaluation of teaching performance*
By Dylan Wiliam (September 2014)

No. 136 *Meeting the challenge of 21st Century schooling*
By Vic Zbar (July 2014)

No. 135 *Assessment: Getting to the essence*
By Geoff N Masters (April 2014)

No. 134 *Education with a Capital E™*
By Charles Fadel (February 2014)

No. 133 *Our Chosen Future: One school's learning reform blueprint*
By Elisabeth Lenders and Liam King (November 2013)

No. 132 *Generating whole-school improvement: The stages of sustained success*
By Vic Zbar (September 2013)

No. 131 *An 'Intercultural understanding' view of the Asia priority: Implications for the Australian Curriculum*
By Eeqbal Hassim (July 2013)

No. 130 *Performance and Development as a driver of teacher and school improvement: Lessons from the field*
By Graham Marshall and Vic Zbar (April 2013)

No. 129 *Culturally responsive schooling*
By Thelma Perso (February 2013)

No. 128 *Developing creativity in students*
By Tim Hawkes (November 2012)

No. 127 *'Cultural Competence' and National Professional Standards for Teachers*
By Thelma Perso (September 2012)

No. 126 *Leading change, changing leadership*
By Patricia Collarbone (July 2012)

No. 125 *Autonomous school leadership, school improvement and the role of professional associations: The importance of 'telling the story'*
By Allan Shaw (April 2012)

Other publications

Leading the education debate Volume 3: Selected papers from the CSE's Seminar Series and Occasional Papers, 2007–2010 (2011)
Editors Vic Zbar and Tony Mackay

This third collection from the CSE Seminar Series and Occasional Papers has contributions from a number of significant contemporary and international education writers. It comprises 15 papers on school improvement and reform, and is organised in five parts: The challenge of implementation; Leadership remains key; Improving classroom practice; Disciplined innovation for improved systems and schools; and A system that engages educators, students and the community.

Volume 3 of *Leading the education debate* continues the sequence of volumes of collections of CSE papers. The two earlier volumes by the same editors, *Leading the education debate* (2003) and *Leading the education debate Volume 2* (2007), are also available from CSE.

Women in school leadership: Journeys to success (2010)
Compiled by Jim Watterston

Twelve women reflect on their personal and professional journeys to school leadership, the barriers they have overcome, the successes they have achieved and what they have learned along the way. Their experiences and advice provide inspiration for any teacher who might aspire to school leadership.

A complete back catalogue of the CSE/IARTV Seminar and Occasional Paper Series, subscription rates to both of these series and more detailed information on any of the publications listed are available on the Centre for Strategic Education's website www.cse.edu.au. Alternatively contact Centre for Strategic Education, phone (+61 3) 9654 1200, fax (+61 3) 9650 5396, email office@cse.edu.au.

Papers in this CSE/IARTV series are intended to encourage discussion of major issues in education. Views expressed by the authors do not necessarily represent views of the Centre for Strategic Education. Comments on papers are most welcome.

The Centre for Strategic Education (CSE) is the business name adopted in 2006 for the Incorporated Association of Registered Teachers of Victoria (IARTV). Therefore, publications which were previously published in the name of IARTV are now published in the name of CSE.

The Centre for Strategic Education welcomes usage of this publication within the restraints imposed by the Copyright Act. Where the material is to be sold for profit then written authority must first be obtained.

The constituent bodies of CSE/IARTV are the Association of Heads of Independent Schools of Australia (Vic) and the Victorian Independent Education Union.

ISSN 1838-8566
ISBN 978-1-921823-60-2

CSE CENTRE FOR STRATEGIC EDUCATION
Leading educational thinking and practice

Editorial Team: Tony Mackay, Keith Redman, Murray Cropley, Andrew Miller