# What makes great teaching?

## Review of the underpinning research

Robert Coe, Cesare Aloisi, Steve Higgins and Lee Elliot Major

**October 2014**

CEM Centre for Evaluation & Monitoring
Durham University
The Sutton Trust

### A framework for professional learning

This review set out to address three apparently simple questions:

- *What makes 'great teaching'?*
- *What kinds of frameworks or tools could help us to capture it?*
- *How could this promote better learning?*

## Question 1: "What makes great teaching?"

### Great teaching is defined as that which leads to improved student progress

We define effective teaching as that which leads to improved student achievement using outcomes that matter to their future success. Defining effective teaching is not easy. The research keeps coming back to this critical point: student progress is the yardstick by which teacher quality should be assessed. Ultimately, for a judgement about whether teaching is effective, to be seen as trustworthy, it must be checked against the progress being made by students.

### The six components of great teaching

Schools currently use a number of frameworks that describe the core elements of effective teaching. The problem is that these attributes are so broadly defined that they can be open to wide and different interpretation whether high quality teaching has been observed in the classroom. It is important to understand these limitations when making assessments about teaching quality.

Below we list the six common components suggested by research that teachers should consider when assessing teaching quality. We list these approaches, skills and knowledge in order of how strong the evidence is in showing that focusing on them can improve student outcomes. This should be seen as offering a 'starter kit' for thinking about effective pedagogy. Good quality teaching will likely involve a combination of these attributes manifested at different times; the very best teachers are those that demonstrate all of these features.

### 1. (Pedagogical) content knowledge (Strong evidence of impact on student outcomes)

The most effective teachers have deep knowledge of the subjects they teach, and when teachers' knowledge falls below a certain level it is a significant impediment to students' learning. As well as a strong understanding of the material being taught, teachers must also understand the ways students think about the content, be able to evaluate the thinking behind students' own methods, and identify students' common misconceptions.

### 2. Quality of instruction (Strong evidence of impact on student outcomes)

Includes elements such as effective questioning and use of assessment by teachers. Specific practices, like reviewing previous learning, providing model responses for students, giving adequate time for practice to embed skills securely

and progressively introducing new learning (scaffolding) are also elements of high quality instruction.

**3. Classroom climate (Moderate evidence of impact on student outcomes)**

Covers quality of interactions between teachers and students, and teacher expectations: the need to create a classroom that is constantly demanding more, but still recognising students' self-worth. It also involves attributing student success to effort rather than ability and valuing resilience to failure (grit).

**4. Classroom management (Moderate evidence of impact on student outcomes)**

A teacher's abilities to make efficient use of lesson time, to coordinate classroom resources and space, and to manage students' behaviour with clear rules that are consistently enforced, are all relevant to maximising the learning that can take place. These environmental factors are necessary for good learning rather than its direct components.

**5. Teacher beliefs (Some evidence of impact on student outcomes)**

Why teachers adopt particular practices, the purposes they aim to achieve, their theories about what learning is and how it happens and their conceptual models of the nature and role of teaching in the learning process all seem to be important.

**6. Professional behaviours (Some evidence of impact on student outcomes)**

Behaviours exhibited by teachers such as reflecting on and developing professional practice, participation in professional development, supporting colleagues, and liaising and communicating with parents.

## Question 2: "What kinds of frameworks or tools could help us to capture great teaching?"

**Assessing teacher quality through multiple measures**

A formative teacher evaluation system – based on continuous assessment and feedback rather than a high-stakes test - must incorporate a range of measures, from different sources, using a variety of methods. A key to suitably cautious and critical use of the different methods is to triangulate them against each other. A single source of evidence may suggest the way forward, but when it is confirmed by another independent source it starts to become a credible guide.

Currently available measures can give useful information, but there is a lot of noise around a weak signal, so we must be careful not to over-interpret. If we were to use the best classroom observation ratings, for example, to identify teachers as 'above' or 'below' average and compare this to their impact on student learning we would get it right about 60% of the time, compared with the 50% we would get by just tossing a coin. Therefore, these judgements need to be used with considerable caution.

### Six approaches to teacher assessment

For this review we focused on three approaches to assessing teachers that demonstrate moderate validity in signalling effectiveness:

1. classroom observations by peers, principals or external evaluators
2. 'value-added' models (assessing gains in student achievement)
3. student ratings

Three other approaches had limited evidence:

4. principal (or headteacher) judgement
5. teacher self-reports
6. analysis of classroom artefacts and teacher portfolios

### Classroom observations

Successful teacher observations are primarily used as a formative process – framed as a development tool creating reflective and self-directed teacher learners as opposed to a high stakes evaluation or appraisal. However, while observation is effective when undertaken as a collaborative and collegial exercise among peers, the literature also emphasises the need for challenge in the process – involving, to some extent, principals or external experts.

Levels of reliability that are acceptable for low-stakes purposes can be achieved by the use of high-quality observation protocols. These include using observers who have been specifically trained – with ongoing quality assurance, and pooling the results of observations by multiple observers of multiple lessons.

### Measuring student gains

Value-added models are highly dependent on the availability of good outcome measures. Their results can be quite sensitive to some essentially arbitrary choices about which variables to include and what assumptions underpin the models. Estimates of effectiveness for individual teachers are only moderately stable from year to year and class to class. However, it does seem that at least part of what is captured by value-added estimates reflects the genuine impact of a teacher on students' learning.

*Student ratings*

Collecting student ratings should be a cheap and easy source of good feedback about teaching behaviours from a range of observers who can draw on experience of many lessons. There is evidence of the validity of these measures from use both in schools and, more widely, in higher education.

## Question 3: "How could this promote better learning?"

A review by Timperley et al. details a teacher 'knowledge-building cycle ' - a feedback loop for teachers – that is associated with improved student outcomes. Their synthesis 'assumes that what goes on in the black box of teacher learning is

fundamentally similar to student learning'. And their findings suggest that teacher learning can have a sizeable impact on student outcomes.

The observation/feedback routine should be structured explicitly as a continuous professional learning opportunity that enables them to work on improving student outcomes.

The literature provides a challenge to the much quoted claim that teachers typically improve over their first 3-5 years and then plateau. Teachers working in schools with more supportive professional environments continued to improve significantly after three years, while teachers in the least supportive schools actually declined in their effectiveness. Another study found that feedback from classroom observation led to a gain in students' math test scores in the years following the intervention, equivalent to an effect size of 0.11.

**Six principles of teacher feedback**

Sustained professional learning is most likely to result when:

1. the focus is kept clearly on improving student outcomes;
2. feedback is related to clear, specific and challenging goals for the recipient;
3. attention is on the learning rather than to the person or to comparisons with others;
4. teachers are encouraged to be continual independent learners;
5. feedback is mediated by a mentor in an environment of trust and support;
6. an environment of professional learning and support is promoted by the school's leadership.

## Contents

## Introduction

This paper sets out to address some apparently simple questions:

- *What is good pedagogy?*
- *What kinds of frameworks or tools could help us to capture it?*
- *How could this promote better learning?*

In focusing on these questions, we recognise that it may seem more obvious to start thinking about teachers' professional learning and development by focusing on the necessary conditions for such learning to occur. For example, we might argue that teachers need to feel trusted and valued, that their experiences and perspectives are acknowledged, that the culture of the schools in which they work should promote critical questioning and innovative approaches, with space and encouragement for discussion and sharing of ideas. We will return to these issues, but first we focus on what that learning should be. Again, it might seem obvious that this is already well known: we surely know what great teaching looks like; we just need to create the culture in which teachers feel empowered and free to do it.

In fact, there is some evidence that an understanding of what constitutes effective pedagogy – the method and practice of teaching – may not be so widely shared, and even where it is widely shared it may not actually be right (Strong et al, 2011; Hamre et al, 2009). Hence it is necessary to clarify what is known about effective pedagogy before we can think about how to promote it. Unless we do that there is a real danger that we end up promoting teaching practices that are no more – and perhaps less – effective than those currently used.

We also review research that has shed some light on what works in terms of the practices of professional learning – whether it is the frameworks used to define teaching effectiveness or observing peers, entering into dialogue and feedback and helping to improve practice.

This study presents a brief review of the existing research evidence that is relevant to these questions. The original research questions we set out to address are given in full in Appendix A.

# What is good pedagogy? Elements of teaching effectiveness

## Defining 'good pedagogy'

Defining effective teaching is of course problematic. Ideally, we might define effective teaching as that which leads to high achievement by students in valued outcomes, other things being equal. We acknowledge that available assessments – and particularly those that have been used for high-stakes accountability or in existing research studies – may not fully capture the range of the outcomes that we might specify as desirable aims for education (Popham and Ryan, 2012; Muijs et al, 2014; Polikoff, 2014).

We also acknowledge that 'other things being equal' may be open to different interpretations about what factors should or can be taken into account. A number of factors will influence students' achievements, for example, pre-existing student characteristics (both of individual students and collectively), characteristics of the school and of the teacher (some of which may be alterable, others not), and of the context. In practice, the attribution of an 'effect' to an individual teacher or school is generally determined by what cannot be explained by factors that are judged to be outside the control of that individual (Raudenbush, 2004). This kind of 'residual attribution' – interpreting value-added simplistically as the effect of the teacher – is, of course, problematic (Newton et al, 2010; Hill et al, 2011; Dumay et al, 2013).

Despite these limitations, wherever possible, it makes sense to judge the effectiveness of teaching from its impact on assessed learning. If the assessments and value-added models available to us are not good enough, we need to improve them. In the meantime we must exercise some caution in interpreting any claims about teaching effectiveness.

A further concern is that in practice, any kinds of observational measures provide at best poor approximations to how much students actually learn. Whether they are based on classroom observation, student surveys, book scrutiny or other sources, their predictive power is usually not high. For example, even in a high-quality research study such as the Measures of Effective Teaching Project (Mihaly et al, 2013, Table 3, p24), the median correlation between a range of value-added and observation ratings was only 0.3. Although a correlation of 0.3 will often be presented as 'highly significant' by researchers, in practice it means that if we were to use classroom observation ratings to identify teachers as 'above' or 'below' average in their impact on student learning we would get it right about 60% of the time, compared with the 50% we would get by just tossing a coin. It is better than chance, but not by much; there is information in classroom observation, but not enough to base important decisions on it. And of course, this is a best-case: with regular teachers or principals using un-validated observation protocols and no quality assurance process to check judgements are aligned, the correlation will be much less, perhaps even negative (Strong et al, 2011).

## Developing indicators of good pedagogy that can be used reliably

There are at least two kinds of problems we could encounter in trying to 'operationalise' good pedagogy - that is developing a set of measures of good (and great) pedagogy that can be reliably used to assess teacher effectiveness. One is to be too specific: to define it in terms of a checklist of observable, effective practices or skills. A potential problem with trying to reduce great teaching to constituent elements is that the whole may be greater than the sum of its parts. The choices a teacher makes in orchestrating their skills may be an essential part of what makes them effective. Focusing on the behaviours themselves will always be too limited. Instead we need to think in terms of a professional pedagogy in which judgement is an essential component. Nevertheless, evaluating the quality of such choices is unlikely to be straightforward.

The other problem is not to be specific enough. Although it is important to be clear about the principles that underpin pedagogy (James and Pollard, 2011), we must also relate them to something that is observable. Theory must be specific enough to be empirically testable and a guide to well-defined actions.

Shulman (1988, p38) has written of the need for "a union of insufficiencies, a marriage of complements, in which the flaws of individual approaches to assessment are offset by the virtues of their fellows". His argument was that although each individual measure of some aspect of teaching effectiveness may be flawed and inadequate, when our view is informed by a varied collection of such measures their failings can be overcome. However, this view seems not to take into account how we might assess the teacher's role in selecting and orchestrating these 'effective' approaches, nor does it address the practical difficulties of turning an array of insufficient indicators into a meaningful whole. Indeed, Shulman himself seems later to have retracted this view (Shulman, 2009). Before we can think about the validity of any measures of teaching effectiveness we need to be clear what those measures are intended to be used for. On some wish-lists will be requirements: for use in selection for initial professional entry; for awarding certification as a qualified teacher; for recognising professional progression, perhaps linked to probation, tenure, promotion, retention, or performance-related pay; for identifying under-performing teachers, with associated support or firing. Unfortunately, the evidence seems clear that our best currently available measures of teaching effectiveness are not adequate for most of these kinds of purposes (Gitomer, 2009).

Our purpose here is a little different. We take the view that low-stakes, formative use of teaching effectiveness indicators, with an emphasis on feedback, support and challenge, and professional learning, may lead to improvements in student learning, even if those indicators are in many ways 'insufficient'. In this we echo Shulman's (2009) distinction between assessment *of* teaching and assessment *for* teaching. However, where Shulman emphasises creating measures for which 'the very act of preparing for and engaging in assessment would be a powerful form of professional development' (p241), we also stress the role of feedback from and discussion about the results of an assessment in professional learning, and the

role of a clearly specified framework of performance indicators to focus teachers' attention and effort on things that are important.

With this approach, our criterion for validating a measure of teaching effectiveness is not 'Does it produce a complete, unbiased and accurate measure of a teacher's impact on student learning?', but 'Can using it as part of a system of self-evaluation, feedback, dialogue and re-assessment lead to improvements in student learning?'. In technical terms, we value consequential validity over criterion-related validity. This perspective also allows us to acknowledge that quality teaching is multidimensional: a profile of multiple, independent strengths and weaknesses may be more useful – and a better fit to reality – than a single, unidimensional measure.

## Types of evidence relevant to 'effectiveness'

There are a number of sources of evidence about the skills, knowledge, behaviours, qualities and competences required to be an excellent teacher. A key feature of the current review is that we try to limit our attention to well-defined, operationalisable behaviours, skills or knowledge that have been found to be related, with at least some justification for a causal relationship, to measureable, enhanced student outcomes. Following Rosenshine (2010, 2012) and Muijs et al (2014), these sources of evidence include:

- Evidence from educational effectiveness research about teacher behaviours associated with learning gains

- Evidence from intervention studies about what can be changed, and its effect on outcomes

- Evidence and theory from cognitive science about learning: how our brains acquire, make sense of and use information

There are two key requirements for the inclusion of a teaching approach as 'great teaching' in this review:

- There must be a clear, well-specified and implementable intervention associated with promoting the approach. It has to be something we can change. For example, the knowledge that 'great teachers have high expectations' is of no use to us unless we have a strategy for encouraging teachers to raise their expectations

- There must be some evidence linking the approach with enhanced student outcomes. There is not necessarily any assumption that such outcomes should be limited to academic attainment: whatever is valued in education should count.

One of the features of research on effective practices is that there are a number of reviews available with quite different claims about what characteristics of teacher practice are associated with improved outcomes. For example, a review by Husbands and Pearce (2012) contains 'Nine claims from research', of which the first is that 'Effective pedagogies give serious consideration to pupil voice' (p3). A

good definition of ‘pupil voice’ is given, but as far as we can tell, none of the studies cited contain robust evidence to link it causally to improvements in pupil outcomes. There is some evidence of a link to changes in teachers’ practices and perceptions, and to more positive attitudes for both teachers and students, though many of even these studies would not meet basic quality standards for robust support of such claims. Using pupil voice may indeed be an effective pedagogy, but we believe that the evidence currently available does not support this claim, so have not included it.

However, we acknowledge that the question of what teaching practices are shown by research to be effective remains contested. An example from England is Brown et al’s (2001) analysis of different views of the research basis of the National Numeracy Strategy. From the US an example is Boaler’s (2008) critique of *The Final Report of the National Mathematics Advisory Panel* (National Mathematics Advisory Panel, 2008).

It is also clear that a lot of the research which has set out to discover the elements of effective teaching have simply asked the wrong questions. As Good and Biddle pointed out more than 25 years ago, looking back then over at least 20 years of this kind of research,

> At various times educators in this century have advocated as answers large-group instruction, smallgroup teaching and individualised teaching!...However it seems clear that simple characteristics of instruction have never predicted instructional effectiveness...The issue is not individualised instruction or small-group instruction, but rather the quality of thought and effort that can occur within these structures...(Good & Biddle, 1988 p.116)

A salutary example is from Brown et al (2001), who confidently identified a list of instructional practices that empirically distinguished effective from less effective teachers, as determined by their students’ learning gains. They then tested the predictive power of an observation schedule based on evaluating these practices for a different group, but found the results rather disappointing:

> We are therefore left with the perhaps rather happy conclusion that the behaviour of effective teachers and less effective teachers are not easily characterised; much depends on the particular way that teachers and classes as people relate together. There are signs that certain types of behaviour may often lead to higher gains, but there are always exceptions in both directions.

A final caution is from the US *National Mathematics Advisory Panel* (2008):

> Unfortunately, little is known from existing high-quality research about what effective teachers do to generate greater gains in student learning. Further research is needed to identify and more carefully define the skills and practices underlying these differences in teachers’ effectiveness, and how to develop them in teacher preparation programs.

## Examples of effective practices

In this section we present a collection of teacher behaviours, approaches, classroom practices and skills that meet our criteria of being well-defined, implementable and linked to gains in student outcomes. We have sought to include here some practices that are counterintuitive, or that challenge the accepted orthodoxy about what is effective teaching, on the grounds that these examples may have value more as a prompt to critical questioning rather than a checklist of desirable behaviours. Teachers may need to have clear understanding of why, when and how each of these practices can be effective, and exactly what it means to demonstrate them in a way that is optimal to promote students' learning. Good summaries of the wider evidence about effective practices can be found in Muijs et al (2014) and in Ko et al (2013).

Some important caveats are required before presenting these examples of 'effective practice'. All of them are open to interpretation. All of them could be done well or done badly. All of them could be inappropriate in some contexts and appropriate in others. For these reasons it may be unproductive or even harmful to treat them as if their meaning is unproblematic or to require them as a recipe or formula. Nevertheless, they are all supported by robust evidence of positive impact on student learning, so may be seen as offering at least a 'starter kit' for thinking about effective pedagogy.

### Danielson's Framework for Teaching

The use of this framework as a classroom observation instrument is discussed in more detail below (p31), but for now we present an outline of the elements that are evaluated.

1. **Planning and preparation**
    a. Demonstrating Knowledge of Content and Pedagogy
    b. Demonstrating Knowledge of Students
    c. Setting Instructional Outcomes
    d. Demonstrating Knowledge of Resources
    e. Designing Coherent Instruction
    f. Designing Student Assessments

2. **Classroom environment**
    a. Creating an Environment of Respect and Rapport
    b. Establishing a Culture for Learning
    c. Managing Classroom Procedures
    d. Managing Student Behaviour
    e. Organizing Physical Space
3. **Instruction**
    a. Communicating with Students
    b. Using Questioning and Discussion Techniques
    c. Engaging Students in Learning
    d. Using Assessment in Instruction
    e. Demonstrating Flexibility and Responsiveness
4. **Professional responsibilities**
    a. Reflecting on Teaching

b. Maintaining Accurate Records
c. Communicating with Families
d. Participating in the Professional Community
e. Growing and Developing Professionally
f. Showing Professionalism

## The Classroom Assessment Scoring System (CLASS)

CLASS (Pianta, La Paro, & Hamre, 2008) is an evaluation framework for classroom observation that identifies three main domains and a number of dimensions within each:

### Emotional Support

Classroom climate (positive and negative) – warmth, respect, enjoyment, enthusiasm
Teacher sensitivity to student needs
Regard for student perspectives – respect for student autonomy, interests, motivations

### Classroom Organization

Behavior management
Productivity – time management, maximizing opportunity to learn
Instructional learning formats – activities that maximize engagement

### Instructional Support

Concept development – focus on higher order thinking
Quality of feedback
Language modelling – questioning, expanding, use of vocabulary

## Rosenshine's Principles of Instruction

Rosenshine (2010, 2012) has summarised at least 40 years of research on effective instruction with a key set of principles that maximise its impact. The starting point for this evidence base is a set of correlational studies linking particular observed classroom teacher behaviours with higher student outcomes. For each of these principles there is also experimental evidence showing that attempts to train teachers in adopting these behaviours can result in changes in teacher behaviours and improvements in student outcomes.

In outline the ten principles are:

1. Begin a lesson with a short review of previous learning
2. Present new material in small steps, with student practice after each step
3. Ask a large number of questions and check the responses of all students
4. Provide models for problem solving and worked examples
5. Guide student practice

6. Check for student understanding
7. Obtain a high success rate
8. Provide scaffolds for difficult tasks
9. Require and monitor independent practice
10. Engage students in weekly and monthly review

### Creemers and Kyriakides' Dynamic Model

A huge body of research in the educational effectiveness tradition has focused on the characteristics of schools and teachers that are associated with high learning gains. Much of the evidence is correlational, cross-sectional and lacking a strong theoretical foundation (Scheerens et al, 2001). However, the Dynamic Model (Creemers & Kyriakides, 2006, 2011) is empirically grounded, well enough specified to be testable and has indeed been subjected to considerable testing and verification.

The model identifies 21 particular teaching practices, grouped under eight headings. Creemers & Kyriakides (2011) have also developed a set of instruments for capturing these practices, consisting of two low-inference classroom observation instruments, a high-inference observational instrument and a student questionnaire, together with a teacher questionnaire for measuring school factors.

**Table 1: The dynamic model of educational effectiveness (Creemers & Kyriakides, 2006)**

| | |
|---|---|
| **(1) Orientation** | **(a) Providing the objectives for which a specific task/lesson/series of lessons take(s) place**<br>**(b) Challenging students to identify the reason why an activity is taking place in the lesson.** |
| **(2) Structuring** | (a) Beginning with overviews and/or review of objectives<br>(b) Outlining the content to be covered and signalling transitions between lesson parts<br>(c) Drawing attention to and reviewing main ideas. |
| **(3) Questioning** | (a) Raising different types of questions (i.e., process and product) at appropriate difficulty level<br>(b) Giving time for students to respond<br>(c) Dealing with student responses. |
| **(4) Teaching modelling**<br>**.** | (a) Encouraging students to use problem-solving strategies presented by the teacher or other classmates<br>(b) Inviting students to develop strategies<br>(c) Promoting the idea of modelling |
| **(5) Application** | (a) Using seatwork or small-group tasks in order to provide needed practice and application opportunities<br>(b) Using application tasks as starting points for the next step of teaching and learning. |
| **(6) The classroom as a learning environment** | (a) Establishing on-task behaviour through the interactions they promote (i.e., teacher–student and student–student interactions)<br>(b) Dealing with classroom disorder and student competition through establishing rules, persuading students to respect them and using the rules. |
| **(7) Management of time** | (a) Organizing the classroom environment<br>(b) Maximizing engagement rates. |
| **(8) Assessment** | (a) Using appropriate techniques to collect data on student knowledge and skills<br>(b) Analysing data in order to identify student needs and report the results to students and parents.<br>(c) Teachers evaluating their own practices. |

## Evidence from cognitive psychology

Because of the fragmentation of academic disciplines, a parallel source of evidence can be found in research in cognitive psychology that has investigated the nature of learning, the conditions under which it occurs and the role of memory in this process. A good summary can be found in Bransford, Brown, & Cocking (2000).

One paradoxical finding is that some approaches that may appear to make learning harder in the short term, and less satisfying for learners, actually result in better long-term retention. Emphasising the difference between short-term *performance* and long-term *learning*, Bjork and Bjork (2011) call these 'desirable difficulties', and give four specific examples:

- **Varying the Conditions of Practice**: Varying the learning context, types of task or practice, rather than keeping them constant and predictable, improves later retention, even though it makes learning harder in the short term.

- **Spacing Study or Practice Sessions**: The same amount of time spent reviewing or practising leads to much greater long-term retention if it is spread out, with gaps in between to allow forgetting. This "is one of the most general and robust effects from across the entire history of experimental research on learning and memory." (Bjork and Bjork, 2011, p59).

- **Interleaving versus Blocking Instruction on Separate To-Be-Learned Tasks**: Learning in a single block can create better immediate performance and higher confidence, but interleaving with other tasks or topics leads to better long-term retention and transfer of skills.

- **Generation Effects and Using Tests (Rather Than Presentations) as Learning Events**: Having to generate an answer or procedure, or having to retrieve information – even if no feedback is given – leads to better long-term recall than simply studying, though not necessarily in the short-term. Testing can also support self-monitoring and focus subsequent study more effectively. "Basically, any time that you, as a learner, look up an answer or have somebody tell or show you something that you could, drawing on current cues and your past knowledge, generate instead, you rob yourself of a powerful learning opportunity" (Bjork and Bjork, 2011, p61).

A recent and comprehensive summary of the impact, strength of evidence and generality of conditions under which a number of learning techniques have been shown to be effective is presented by Dunlosky et al (2013).

**Table 2: Effectiveness of ten learning techniques, from Dunlosky et al (2013)**

| | |
|---|---|
| High utility | **Practice testing** |
| | Self-testing or taking practice tests on material to be learned |
| | **Distributed ('spaced') practice** |
| | Implementing a schedule of practice that spreads out study activities over time |
| Moderate utility | **Elaborative interrogation** |
| | Generating an explanation for why an explicitly stated fact or concept is true |
| | **Self-explanation** |
| | Explaining how new information is related to known information, or explaining steps taken during problem solving |
| | **Interleaved practice** |
| | Implementing a schedule of practice that mixes different kinds of problems, or a schedule of study that mixes different kinds of material, within a single study session |
| | **Summarization** |
| | Writing summaries (of various lengths) of to-be-learned texts |
| Low utility | **Highlighting** |
| | Marking potentially important portions of to-be-learned materials while reading |
| | **Keyword mnemonic** |
| | Using keywords and mental imagery to associate verbal materials |
| | **Imagery use for text learning** |
| | Attempting to form mental images of text materials while reading or listening |
| | **Rereading** |
| | Restudying text material again after an initial reading |

## Examples of teacher characteristics

As well as observable behaviours, there are also some teacher characteristics that may not be directly observable in classroom behaviour, but which have been found to be related to students' learning gains.

### (Pedagogical) Content knowledge

A number of studies have found a relationship between measures of a teacher's knowledge of the content they are teaching and the gains made by their students. It seems intuitively obvious that 'Teachers cannot help children learn things they themselves do not understand' (Ball, 1991, p5). However, the search for a relationship between characteristics such as academic qualifications or general ability and student performance has been rather disappointing: correlations are typically very small or non-existent (Rockoff et al, 2011). Nevertheless, there seems to be an emerging body of work that can link more specific measures of

content knowledge, and in particular the kinds of content knowledge that are relevant to teaching, to student gains.

For example, Sadler et al (2013) tested a group of volunteer, experienced middle school (seventh and eighth grade) science teachers on their understanding of the content they were teaching and on the kinds of misconceptions they expected students to show. Generally, their understanding of the content was good, though there was enough variation to give some predictive power to teachers' subject knowledge: overall, teachers answered 83% correctly, compared with 38% by their students. However, the teachers' ability to identify common misconceptions was hardly above chance. Overall, there was a positive but modest relationship between teachers' understandings and their students' gains. However, an item-level analysis of the relationship between teachers' and students' understanding of specific concepts had considerably more predictive power. This suggests that targeting support for teachers at particular areas where their understanding or their knowledge of student misconceptions is weak may be a promising strategy, a claim that is supported by reviews of the impact of teacher professional development in these areas (Timperley et al, 2007; Blank and de las Alas, 2009).

Hill et al (2005) investigated the importance of teachers' pedagogical content knowledge in mathematics. They cited a number of studies that have found that teachers' level of understanding of the mathematics they are teaching is related to how effectively students learn it. In their own analysis, they found that the difference between high and low scoring (a 2 SD gap) teachers on their Content Knowledge for Teaching (CKT) was associated with more than a month's additional learning for students in a year. Although this is not a huge effect, it is of similar order to the strength of the relationship between socioeconomic background and attainment, for example. Interestingly, most of the difference was between the lowest scoring teachers and the rest: once their CKT score was into the third decile there was no further relationship with student learning.

### Beliefs about learning

Askew et al (1997) found that highly effective teachers of numeracy were characterised by a particular set of beliefs, which in turn led to a corresponding set of teaching approaches. They claim that "The mathematical and pedagogical purposes behind particular classroom practices are as important as the practices themselves in determining effectiveness" (p5). In other words, simply describing or defining observable practices or approaches is not enough to characterise teachers as more or less effective; it matters *why* the teachers adopt them.

In particular, Askew et al (1997) identified beliefs about the nature of mathematics and what it means to understand it, along with teachers' beliefs and theory about how children learn and about the teacher's role in promoting learning, as important distinguishing factors between those who were more and less effective (see table 3). Given the potential significance of the need to focus on teacher beliefs, it seems surprising that these findings do not seem to have been extensively tested by further research; although there is extensive research on teacher beliefs, links with pupil progress are much less common. A study by Higgins and Moseley (2001) of teacher beliefs about Information and

Communication Technology failed to find any convincing relationships between beliefs and pupil progress.

However, some corroboration can be found in the evidence from Timperley et al (2007) that the professional development programmes with demonstrable benefits for learners mostly included some attempt to engage with teachers' existing theories, values and beliefs (p196). Such a claim is also consistent with a view of effective pedagogy as consisting of more than just a set of classroom techniques, but depending on the ability to make complex judgements about which technique to use when.

**Table 3: Characteristics of highly effective teachers of numeracy, from Askew et al, 1997**

**Highly effective teachers were characterised by beliefs about**

***What it means to be numerate:***

- having a rich network of connections between different mathematical ideas
- being able to select and use strategies, which are both efficient and effective.

  They used corresponding teaching approaches that:
  - connected different areas of mathematics and different ideas in the same area of mathematics using a variety of words, symbols and diagrams
  - used pupils' descriptions of their methods and their reasoning to help establish and emphasise connections and address misconceptions
  - emphasised the importance of using mental, written, part-written or electronic methods of calculation that are the most efficient for the problem in hand
  - particularly emphasised the development of mental skills.

***How children learn:***

- almost all pupils are able to become numerate
- pupils develop strategies and networks of ideas by being challenged to think, through explaining, listening and problem solving.

  They used teaching approaches that:
  - ensured that all pupils were being challenged and stretched, not just those who were more able
  - built upon pupils' own mental strategies for calculating, and helped them to become more efficient.

***The role of the teacher:***

- discussion of concepts and images is important in exemplifying the teacher's network of knowledge and skills and in revealing pupils' thinking
- it is the teacher's responsibility to intervene to assist the pupil to become more efficient in the use of calculating strategies.

  These teachers used teaching approaches that encouraged discussion, in whole classes, small groups, or with individual pupils.

**Less effective teachers believed in the importance of either**

- pupils acquiring a collection of facts and standard methods, and that pupils varied in their ability to remember these. They used teaching approaches that:
  - dealt with areas of mathematics discretely
  - emphasised teaching and practising standard methods and applying these to abstract or word problems without considering whether there were alternative more efficient ways of solving a particular problem.

or

- developing numeracy concepts using practical equipment and waiting until pupils were ready to move onto more formal methods. They used teaching approaches that emphasised pupils working things out for themselves, using any method with which they felt comfortable.

## Other characteristics

A large number of studies have set out to find links between a variety of other teacher characteristics and student achievement gains. Wayne and Youngs (2003) conducted a review of the available literature and concluded that there were positive (though often inconsistent and probably small) associations between

student learning gains and teacher characteristics such as the status of the college they had attended or their scores on certain kinds of tests, such as licensure or reasoning tests, or specific tests of the material they were teaching.

For mathematics teachers, having a higher degree in maths, or a better class of degree, was associated with more student learning, but the same relationship was not found in other subjects. Similarly, being certified (qualified) in maths or science teaching was associated with greater effectiveness, but there was no relationship between certification and effectiveness in other subjects. Ball and Hill (2009) review some of the later literature on the relationships between teacher certification, qualifications and level of study with student learning, and conclude they are generally inconsistent and hard to interpret.

Interestingly, a number of teacher characteristics (such as teachers' self-reported self-efficacy, extraversion and conscientiousness) were found by Rockoff et al (2011) to be related to supervisor ratings of effectiveness but not to actual student achievement gains.

## Examples of ineffective practices

It may seem unduly negative to focus on things that do not work, but there are a number of reasons for wanting to do this.

One is that it provides a challenge to complacency. A potential problem with lists of 'best practice' is that they can be susceptible to confirmation bias. If the list of effective practices is long enough, and contains descriptions of practices that are open to a bit of interpretation, most teachers will be able to identify some they think they are doing. Such lists can also seem, like motherhood and apple pie, to be good, but predictable, obvious and nothing new. Including some examples of 'worst practice' is likely to provoke a stronger reaction, which we hope can be challenging in a constructive way. Clearly, bluntly telling a teacher that some aspect of their practice is wrong may not be a good way to get a discussion going, however.

A second reason is that many of these ineffective practices seem to be quite popular, though most evidence here is anecdotal and selective. It may be that as well as telling us 'what works', an important contribution of research is to tell us what doesn't work. By stopping doing things that are either ineffective or inefficient, we should allow more time to focus on thing that will make more difference.

The following are examples of practices whose use is not supported by research evidence:

*Use praise lavishly*

Praise for students may be seen as affirming and positive, but a number of studies suggest that the wrong kinds of praise can be very harmful to learning. For example, Dweck (1999), Hattie & Timperley (2007).

Stipek (2010) argues that praise that is meant to be encouraging and protective of low attaining students actually conveys a message of the teacher's low

expectations. Children whose failure was responded to with sympathy were more likely to attribute their failure to lack of ability than those who were presented with anger.

> "Praise for successful performance on an easy task can be interpreted by a student as evidence that the teacher has a low perception of his or her ability. As a consequence, it can actually lower rather than enhance self-confidence. Criticism following poor performance can, under some circumstances, be interpreted as an indication of the teacher's high perception of the student's ability." (ibid)

*Allow learners to discover key ideas for themselves*

Enthusiasm for 'discovery learning' is not supported by research evidence, which broadly favours direct instruction (Kirschner et al, 2006). Although learners do need to build new understanding on what they already know, if teachers want them to learn new ideas, knowledge or methods they need to teach them directly.

*Group learners by ability*

Evidence on the effects of grouping by ability, either by allocating students to different classes, or to within-class groups, suggests that it makes very little difference to learning outcomes (Higgins et al, 2014). Although ability grouping can in theory allow teachers to target a narrower range of pace and content of lessons, it can also create an exaggerated sense of within-group homogeneity and between-group heterogeneity in the teacher's mind (Stipek, 2010). This can result in teachers failing to make necessary accommodations for the range of different needs within a supposedly homogeneous 'ability' group, and over-doing their accommodations for different groups, going too fast with the high-ability groups and too slow with the low.

*Encourage re-reading and highlighting to memorise key ideas*

This finding has already been mentioned in summarising the review by Dunlosky et al (2013). Re-reading and highlighting are among the commonest and apparently most obvious ways to memorise or revise material. They also give a satisfying – but deceptive – feeling of fluency and familiarity with the material (Brown et al, 2014). However, a range of studies have shown that testing yourself, trying to generate answers, and deliberately creating intervals between study to allow forgetting, are all more effective approaches.

*Address issues of confidence and low aspirations before you try to teach content*

Teachers who are confronted with the poor motivation and confidence of low attaining students may interpret this as the cause of their low attainment and assume that it is both necessary and possible to address their motivation before attempting to teach them new material. In fact, the evidence shows that attempts to enhance motivation in this way are unlikely to achieve that end. Even if they do, the impact on subsequent learning is close to zero (Gorard, See & Davies, 2012). In fact the poor motivation of low attainers is a logical response to repeated failure. Start getting them to succeed and their motivation and confidence should increase.

*Present information to learners in their preferred learning style*

A belief in the importance of learning styles seems persistent, despite the prominence of critiques of this kind of advice. A recent survey found that over 90% of teachers in several countries (including the UK) agreed with the claim that "Individuals learn better when they receive information in their preferred learning style (for example, visual, auditory or kinaesthetic)" (Howard-Jones, 2014). A number of writers have tried to account for its enduring popularity (see, for example, a clear and accessible debunking of the value of learning styles by Riener and Willingham, 2010), but the psychological evidence is clear that there are no benefits for learning from trying to present information to learners in their preferred learning style (Pashler et al, 2008; Geake, 2008; Riener and Willingham, 2010; Howard-Jones, 2014).

*Ensure learners are always active, rather than listening passively, if you want them to remember*

This claim is commonly presented in the form of a 'learning pyramid' which shows precise percentages of material that will be retained when different levels of activity are employed. These percentages have no empirical basis and are pure fiction. Memory is the residue of thought (Willingham, 2008), so if you want students to remember something you have to get them to think about it. This might be achieved by being 'active' or 'passive'.

## How do we measure it? Frameworks for capturing teaching quality

### Section summary

This section reviews the range of different approaches to the evaluation of teaching. Goe, Bell & Little (2008) identify seven methods of evaluation:

- classroom observations, by peers, principals or external evaluators
- 'value-added' models (assessing gains in student achievement)
- student ratings
- principal (or headteacher) judgement
- teacher self-reports
- analysis of classroom artefacts
- teacher portfolios

For this review we define "observation-based assessment" as all measurement activities whose main task is to watch teachers deliver their lesson, whether in real time or afterwards, and regardless of who is carrying out the assessment. We summarise research on observations performed by: teacher colleagues, senior management or principals, external inspectors, students, and self-reports.

### Classroom observation approaches

Classroom observations are the most common source of evidence used in providing feedback to teachers in OECD countries, whether American (e.g. Canada, Chile, United States), European (e.g. Denmark, France, Ireland, Spain) or Asian-Pacific (e.g. Australia, Japan, Korea).

Successful teacher observations are primarily used as a formative process – framed as a development tool creating reflective and self-directed teacher learners as opposed to a high stakes evaluation or appraisal. However, while observation is effective when undertaken as a collaborative and collegial exercise among peers, the literature also emphasises the need for challenge in the process – involving to some extent principals or external experts. It suggests that multiple observations are required using a combination of approaches.

**Evidence** of impact on student outcomes is generally limited. This highlights a common challenge identified throughout the research: while the theoretical principles of observation are uncontroversial among teachers, the actual consistent disciplined implementation is far more difficult. Teachers or head teachers must be trained as observers – otherwise well intentioned programmes can revert to the blind leading the blind.

Another recurring theme in the research is that any successful programme of teacher observation (whether a peer or a principal, from inside or outside the school), needs to address educational and **political** challenges dealing with issues of trust, authority, and knowing who is in charge of the information generated.

## Peer observations

Overall, the research literature presents a positive **narrative about** peer observation as a driver of both teacher learning and a school's sense of collaboration and collegiality. It is primarily effective as a formative process where the teacher observed has full control over what happens to information about their observation.

However its **effective adoption** depends very much on the willingness of all parties involved to contribute. This is a political as well as educational issue**. Evidence** of impact on student outcomes is limited.

*Peer observation as a formative process*

Bernstein (2008) draws from a range of sources to argue that 'class observations should yield formative review only, unless multiple observations by well-prepared observers using standardized protocols are undertaken' because the reliability of observations by unprepared peers is low (*ibid*., p. 50).

Goldberg et al. (2010) survey 88 teachers and administrators and find that most respondents find peer reviews meaningful and valuable 'for their own personal use – to modify and improve their teaching' (Maeda, Sechtem & Scudder, 2009). The observation is deemed to be useful also by the observers, as it has 'forced them to reflect on their own teaching skills and methods' (Goldberg et al., 2010) and has had an impact on their practice, a result obtained also by Kohut, Burnap & Yon (2007).

According to McMahon and colleagues (2007) 'what really matters is whether or not the person being observed has full control over what happens to information about the observation'. Where this does not happen, teachers may be reluctant to be involved in the observation even when the stakes are not necessarily high. A similar view is shared by Chamberlain, D'Artrey & Rowe (2011), who find that formative observation can become a box-ticking exercise when it is imposed on staff and it is separated from a more formalised development system.

In an Australian study, Barnard et al. (2011) make use of 'peer partnership', which are a form of peer observation in which two teachers 'eyewitness [each other's] teaching and learning activities and […] provide supportive and constructive feedback' (*ibid*., p. 436–437, see also Bell, 2005). They find that while the major hurdle against participation was the commitment in terms of effort and time, once this is overcome teachers felt rewarded by the experience and wanted to continue with the project.

*Peer Assistance and Review (PAR)*

One of the best-documented approaches to peer observations in schools is the Peer Assistance and Review (PAR) protocol deployed in some districts in the US. This programme was based on the idea that teaching practice could be improved by using expert teachers as mentors for beginning teachers 'the way doctors mentor interns' (Kahlenberg, 2007).

Goldstein (2007) finds six features that distinguish PAR from other less effective assessments, and especially from principal observations: (1) 'the amount of time spent on evaluation'; (2) the tight relationship between observations, formative feedback and professional development; (3) 'the transparency of the evaluation process'; (4) the involvement of teacher unions in the strategy and the appraisal; (5) the credibility of the evaluation; and (6) 'the degree of accountability' involved in the process.

For this system to work a number of conditions must be in place: there must be agreement from all stakeholders on who the mentors will be and what their role is; there must be agreement on what the teaching standards are and how to measure quality, effectiveness or improvement; there must be the willingness from both teacher union and principals to delegate part of their power to an *ad hoc* panel; there must be a favourable political context and the strength to stand by some radical departures from the norm; and there must be the resources to pay for the programme.

Overall, the benefits of PAR seems to be mostly indirect: by being 'designed for *selective* retention', PAR 'increases the likelihood that students will have the teachers they deserve' (Johnson et al., 2009).

There are reports of school-wide effective interventions that, like PAR, manage to overcome aversion to integrate both a formative and a summative component. For example, Bramschreiber (2012) describes the model in place in a school in Colorado, consisting in: frequent observations by 'master teachers', who train staff around 'either research-based teaching strategies aligned to a schoolwide goal or general best teaching practices'; a 'Campus Crawl', where twice a year all teachers observe peers in the same or another department; and four formal observations, two conducted by the school managers for summative purposes, and two by the master teachers for formative purposes.

**School leader / principal observations**

Isoré (2009) reports that in OECD countries 60% of students are enrolled in schools where observations are carried out by principals, although the individual country figures are highly variable, going from 100% in the United States to 5% of students in Portugal.

Overall, the literature is that the theory underlying this type of observation – building trusting relationships, empowerment, low-stakes and the need of teacher motivation - are not controversial. The real hurdle is that even after a successful protocol is in place there is still a discrepancy between the '**conversational**' aspects of it (the discourses on the importance of feedback, the talks within the

observation conferences) and its actual **outcomes in practice**. The problem is one of implementation.

Much of the research on principal observations has focused on determining the fairness and reliability of their scoring compared to other measures of teacher effectiveness, such as student (value-added) test scores. The research suggests that without using detailed standard-based instruments and receiving appropriate training, principals are not particularly suited for teacher assessment.

Overall, the findings from Levy & William's (2004) review are aligned with those coming from the literature on peer observations, which were reported in the previous section: 'performance appraisals are no longer just about accuracy, but are about much more including development, ownership, input, perceptions of being valued, and being a part of an organizational team'. This has implications for principal training: if employees must feel supported and that their voice matters, training 'could focus on how to deliver feedback in a supportive, participatory way as opposed to or in addition to other more traditional types of training (Pichler, 2012, p. 725).

Formative feedback is never completely separated from summative judgements. After studying a network of charter schools in the United States, Master (2012) reports that formative mid-year evaluations were still strongly associated to end-of-year dismissals or promotions decisions.

*Examples of successful principal observations*

Range, Young & Hvidston (2013) investigate the effect of the 'clinical supervision' model (see Goldhammer, 1969; Cogen, 1973; cited in Range, Young & Hvidston, 2013), which is comprised of a flow of observations followed by pre- and post-observation meetings (conferences). The pre-observation conference is where the modes, scope and aims of the observation are negotiated and where teachers can present the classroom context. On the post-observation conference, the authors note that it should take place in a comfortable setting **no longer than five days after the observation**. Their feedback should be factual, non-threatening, acknowledging of the teacher's strengths, aimed at creating reflective and self-directed teacher learners (see Ovando, 2005, on how to train principals to write constructive feedback, and Ylimaki and Jacobson, 2011, for a general overview on principal preparation).

Overall, Range, Young & Hvidston (2013) agree with Bouchamma (2005) on the positive response of teachers towards the clinical supervision model and find that a trusting relationship, constructive feedback and the discussion about areas of improvement are valued as important by their sample both in the pre- and in the post-observation conference. Moreover, they find differences in the responses of beginning and experienced teachers, which they interpret as evidence in favour of Glickman's (1990) theory of developmental supervision, according to which novice and struggling teachers would benefit from a more directive leadership approach (Range, Young & Hvidston, 2013).

While the clinical supervision model involves an observation and one pre- and post-observation conferences, other authors have explored the effectiveness of

the ‘negotiated assessment’ (Gosling, 2000, in Verberg, Tigelaar & Verloop, 2013), which is characterised by a ‘learning contract’ between the assessor and the assessed containing ‘the negotiated learning goals, learning activities and the evidence to be provided during the assessment procedure’. Despite the stress on formative feedback, peer observation and empowered, self-directing learning and training, their study reported that while the assessment meetings were useful, collecting and discussing about evidence was far less appealing. This suggests once more that in spite of the theory, the implementation of any strategy has to take into account the practical and intellectual burden asked from teachers for it to produce any effect in the classroom.

Tuytens & Devos (2011) argue, after a study on 414 teachers in Belgium, that active supervision, charisma and content knowledge are all significantly associated with teachers perceived to be effective at feedback. The relationship between principal effectiveness, feedback quality and impact is well summarised by a teacher’s critique to the appraisal system he or she was subject to: ‘It is a one-shot observation and has no lasting impact. The only time it is helpful is when you have an administrator that gives really beneficial feedback. This rarely happens’ (Ovando, 2001, p. 226).

O’Pry & Schumacher (2012) evaluate teacher perceptions of a complex standard-based evaluation system such as the Professional Development Appraisal System (PDAS), used in Texas, and find that the leadership actions have a massive implication on whether the system ends up being accepted or rejected:

> Teachers who feel as though they had a principal or appraiser who was knowledgeable about the system; who valued the system; who took time to make them feel supported and prepared for the experience; who was someone with whom they shared a trusting, collegial relationship; who gave them an opportunity to receive valuable and timely feedback; and who guided them through thoughtful reflection on the appraisal results perceived the evaluation experience as a positive, meaningful one. When any of these factors was absent or lacking in the experience of the teacher, the perception of the teacher regarding the process was quite negative as a whole. (*ibid*., p. 339)

**Observation by an external evaluator**

Teachers and principals **say** that feedback from an external evaluator has spurred change in their classroom/school practices, but whether this change is **actual**, **sustained** and **beneficial** is not clear from the research. Moreover, the literature on school inspections is related to another consistent finding of this review: the fact that whenever a third-party observes a teacher practice (whether a peer or a manager, from inside or outside the school), part of the issues with the assessment are not technical, but **political** in nature, as they involve concepts such as trust, authority, territory and power over the information.

In OECD countries, external school inspections are carried out ‘using professional evaluators, regional inspectors, or a district/state/national evaluation department [as well as] independent evaluation consultant[s]’ (Faubert, 2009, p. 14), which means that there is a range of professionals (usually—but not

always—experienced teachers) that can potentially 'invade' a teacher's space. Although the main outcome of classroom observations is to inform school accountability, in fact, in countries such as Germany, Ireland, the United Kingdom and the Czech Republic the external observations can be accompanied by personalised feedback (Faubert, 2009).

A study on 2400 educators in Hong Kong found that teachers (and especially primary school teachers) were much less willing to welcome observers in their classroom than school management or principals were, perhaps because in this context principal observation is more related to summative than formative feedback (Lam, 2001).

A similar study in elementary schools found that while less experienced teachers thought that senior teachers were better assessors than principals, they were no more ready to accept them over principals as observers for formative purposes and preferred principals for summative ones (Chow et al., 2002). The researchers argued that this could be due to the fact that classroom teachers saw principals as a more authoritative figures, and were therefore more willing to accept consequences coming from someone higher in the hierarchy (*ibid*.).

Mangin (2011) argues that one of the challenges faced by external teacher mentors such as those employed by PAR is that on the one hand they try to gain other teachers' trust by de-emphasising and downplaying their expert status, but at the same time they have to ensure that this does not 'undermine others' perceptions of [their] ability to serve as a resource'. Mangin (2011) suggests that a change in the teachers' professional norms is needed to overcome this paradoxical situation, one where both practitioners and external observers are willing to deal with "hard feedback", that is those 'instances where a teacher leader's honest critique of classroom practice is issued even though the critique actively challenges the teacher's preferred practice and may lead the teacher to experience some level of professional discomfort' (Lord, Cress & Miller, 2008, p. 57, quoted in Mangin, 2011, p. 49).

In an evaluation of three external mentoring programmes for science teachers in English secondary schools, Hobson and McIntyre (2013) report that in many instances teachers were unwilling to expose their weaknesses to senior management or even colleagues because of the negative opinion that other professionals could have of their performance. In this case, the external mentors seemed to provide an effective 'relief valve' for teachers (our wording), because of their 'lack of involvement in [the] assessment or appraisal [of the teachers], as well as […] their perceived trustworthiness and non-judgmental nature, and the promise of confidentiality' (Hobson & McIntyre, 2013, p. 355).

Faubert (2009) reports lack of training and support to act upon evaluation results in a meaningful way, negligible or negative effects of external evaluation and accountability on student results, as well as negative effects on teacher motivation.

There are claims that, after certain tensions are released, external evaluation can complement self-evaluation and serve as a tool for school improvement (Whitby, 2010), but a later systematic review provides a more realistic picture. Klerks

(2012) summarises research findings on the effectiveness of school inspections in raising student achievement and changing teacher behaviours. The author reports that the few studies available provide little to no evidence of any direct effect of external evaluation on student achievement or global school improvement.

Ehren et al. (2013) state that 'we do not know how school inspections drive improvement of schools and which types of approaches are most effective and cause the least unintended consequences' (*ibid*., p. 6), and that in those fewer instances where feedback is followed by changes in teacher practice, these rarely involve 'thorough innovation'. What tends to happen, instead, is a 'repetition of content and tasks', the adoption of assessment task formats, or a 'slight [change] in classroom interaction'.

## Instruments for classroom observation

Although a great number of instruments have been developed over several decades to measure what happens in the classroom, these have filtered down to a relatively few that are now widely used – alongside the national teacher standards that countries including England and Australia have produced.

Some of the protocols currently popular include Charlotte Danielson's Framework for Teaching and Robert Pianta's Classroom Assessment Scoring System™ (CLASS™), but other measures of classroom quality exist: the Assessment Profile for Early Childhood Programs (APECP), the Classroom Practices Inventory (CPI), the UTeach Teacher Observation Protocol (UTOP), Fauth's et al. (2014) Teaching Quality Instrument or the Questionnaire on Teacher Interaction (QTI). A number of other observation instruments are described in Ko et al (2013).

### *Danielson's Framework for Teaching*

The Framework for Teaching [(FfT) Danielson, 1996, revised 2007/2011/2014] is a standard-based teacher evaluation system or rather, according to the website, 'a research-based set of components of instruction grounded in a constructivist view of learning and teaching'[1]. The FfT is used to assess four dimensions of teaching: planning and preparation, classroom environment, instruction and professional responsibilities.

The FfT has gained widespread popularity, and although the exact figures are not known, the website reports it having been adopted 'in over 20 states'[2]. It is in many ways one of the gold standard frameworks available being based in part on research. Technically, the FfT is neither an observational instrument nor an observation and feedback protocol, as it only offers a categorisation of certain teaching practices deemed to be conducive to learning. In fact, *The Framework for Teaching Evaluation Instrument* (Danielson, 2014) suggests that evidence should be gathered not only through direct classroom observations, but also through artefacts and principal conferences.

In order for the evaluation instrument to be implemented as intended by the author, the Danielson Group offers a number of paying workshops ranging from

[1] http://danielsongroup.org/framework/
[2] http://danielsongroup.org/charlotte-danielson/

simple training on the use of rubrics to more complicated professional development programmes[3]. This is a non-negligible point for the purposes of this document, not just because of the costs associated with observer training, but also because FfT has been employed in a variety of settings and with different degrees of alignment to its original structure—which in turn makes it difficult to interpret and generalise the studies.

Borman & Kimball (2005) examine the results for 7,000 students in grades 4-6 in Washoe Country, Nevada, where the FfT has been implemented with 'relatively minor changes' and found that the relationship between the FfT and student achievement was rather weak.

In a review of effective measures of teaching, Goe, Bell & Little (2008) confirm the 'wide variation in rater training, rater's relationship with the teacher, the degree of adherence to Danielson's recommendations for use, the use of scores, and the number of observations conducted for each teacher'. Overall, they conclude that:

- The research does not indicate whether modified versions of the instrument perform as well as versions that adhere to Danielson's recommendations (*ibid*., p. 23)
- It is not evident whether the instrument functions differently […] at different grade levels. (*ibid*.)

More accurate research was carried out in recent years, but the results were not too different: as part of the research for the MET project, another modified version of FfT was found to be only modestly correlated with both academic achievement and a range of socio-emotional and non-cognitive outcomes (Kane et al 2013).

Sartain, Stoelinga & Brown (2011) examine the predictive validity of a modified version of the FfT adopted in Chicago public schools and used in the "Excellent in Teaching" pilot study. They find that 'in the classrooms of highly rated teachers, students showed the most growth' (*ibid*., p. 9), which means that there was a positive correlation between teacher ratings on the FfT and their value-added measure. Moreover, the authors found that principals tend to give higher scores to teachers than external observers because they 'intentionally boost their ratings to the highest category to preserve relationships' (*ibid*., p. 41). Overall, the authors' conclusion is worth sharing in full:

> 'Though practitioners and policymakers rightly spend a good deal of time comparing the effectiveness of one rubric over another, a fair and meaningful evaluation hinges on far more than the merits of a particular tool. An observation rubric is simply a tool, one which can be used effectively or ineffectively. Reliability and validity are functions of the users of the tool, as well as of the tool itself. The quality of implementation depends on principal and observer buy-in and capacity, as well as the depth and quality of training and support they receive.
>
> Similarly, an observation tool cannot promote instructional improvement in isolation. A rigorous instructional rubric plays a critical role in defining

[3] http://danielsongroup.org/services/

effective instruction and creating a shared language for teachers and principals to talk about instruction, but it is the conversations themselves that act as the true lever for instructional improvement and teacher development.'

### *CLASS™*

The Classroom Assessment Scoring System™ was developed by Robert Pianta at the University of Virginia, Curry School of Education, Center for Advanced Study of Teaching and Learning (CASTL). Like the FfT, CLASS™ was chosen by the MET project as one of the instruments to measure teacher effectiveness. Unlike the FfT, though, CLASS™ is a stand-alone observational instrument focusing on classroom organisation, teacher-pupil instructional and emotional support. Researchers at CASTL claim that CLASS™ has been used/validated in over 2000[4] or 6000 (CASTL, 2011) classrooms.

Ponitz et al. (2009) found that one dimension of CLASS™ (classroom organisation), was found to be predictive of 172 first graders' reading achievement in a rural area in the southeast of the United States. The MET Project finds with CLASS™ the same significant but weak correlations observed for FfT (Kane et al 2013), and other researchers are even more critical of it, finding that having access to CLASS™ and training did not help observers to rate teachers more accurately (Strong, Gargani & Hacifazlioğlu, 2011).

### *Subject-specific instruments*

The literature shows that content-specific practices tend to have more impact than generic practices on student learning. Therefore, it could be worth at least pairing general measures of teacher effectiveness with some that are content-based such as, for example, the Protocol for Language Arts Teaching Observation (PLATO, see Grossman et al., 2014, for a comparison between PLATO and different value-added models), the Mathematical Quality of Instruction (MQI)[5] and many others, such as the Reformed Teaching Observation Protocol (RTOP, Sawada et al., 2002), the Practices of Science Observation Protocol (P-SOP, Forbes, Biggers & Zambori, 2013), of the Electronic Quality of Inquiry Protocol (EQUIP) for mathematics and science (Marshall, Horton & White, 2009).

## Value-added measures

The use of value-added models (VAMs) have become extremely controversial in recent years, particularly in the US. The prevalence of regular state-wide testing, encouraged by 'Race to the Top', has allowed widespread linking of student test score gains to the individual teachers who taught them, and some instances of teachers losing their jobs as a result.[6] A number of studies have investigated the validity of VAMs as a measure of teaching quality, or to support particular uses. We summarise the main arguments and evidence here.

---

[4] http://curry.virginia.edu/research/centers/castl/class

[5] http://isites.harvard.edu/icb/icb.do?keyword=mqi_training&tabgroupid=icb.tabgroup120173

[6] "School chief dismisses 241 teachers in Washington". *New York Times,* July 3 2010. Available at www.nytimes.com/2010/07/24/education/24teachers.html

Several studies have compared effectiveness estimates from different VAMs and shown that the results can be quite sensitive to different decisions about these issues. Crucially, these decisions are essentially arbitrary, in the sense that there is not a clear *prima facie* or universally agreed correct approach.

For example, different assessments used as the outcome measure will change the rank order of teachers' scores (Papay, 2011; Lockwood et al 2007). Grossman et al (2014) have claimed that the strength of correspondence between value-added and observation measures also depends on the type of assessment used as the outcome in the value-added model, and that correlations are higher with assessment of "more cognitively complex learning outcomes" than with state tests. Although this is true, in neither case are the correlations (0.16 and 0.09, respectively) particularly impressive.

Hill et al (2011) discuss a range of different approaches to which prior characteristics should be statistically controlled for in VAMs. One dilemma, for example, is whether to subtract an overall 'school effect' from the effects that are attributed to individual teachers in that school (for statisticians, this is the issue of whether to include school-level fixed effects). One could argue that an effect that is shared by all classes in a school may well reflect quality of leadership, compositional effects, or unobserved but pre-existing student characteristics, and hence should not be attributed to individual teachers. On the other hand, if all the teachers in a school happen to be good, it might seem unfair to say that is a 'school effect'; and constraining every school to have a zero sum effectively puts teachers in competition against their colleagues. Nevertheless, as Hill et al (2011) show, different US districts and VAMs have taken each side of this debate.

In their own analysis, Hill et al (2011) found that incorporating student-level demographic variables in the model or school fixed effects changed teacher ranks somewhat, but the use of simple gain scores (an alternative approach favoured by some districts) made a big difference (p808). For example, with four different value added models, two-thirds of their sample would be in the top half if they could choose their best score. In another review by McCaffrey et al (2009) a range of different models were found to give different results.

A related issue is whether leaving out important characteristics that have not even been captured could bias the results. Chetty et al (2011) tested teachers' value-added estimates to see whether they were affected by key variables that had not been included in the models and found that there was no evidence of bias. Individual teachers' value-added scores were also consistent across changes from one school to another. They also found long lasting effects on students of being taught by a teacher with high value-added scores, for example being more likely to attend college, earning more money on average and being less likely to become a teenage parent.

Reardon and Raudenbush (2009) set out to examine the assumptions required to interpret value-added estimates of learning gain as a causal effect of teaching. Overall, they conclude that there is considerable sensitivity in these models to a number of assumptions that are either implausible or untestable (or both).

A range of evidence suggests that VAMs can be affected by the effects of prior teachers, measurement error, and the distribution of students into classrooms and teachers into schools (Hill et al, 2011; Amrein-Beardsley, 2008; Kupermintz, 2003; McCaffrey et al., 2003).

Kennedy (2010) points out that our natural tendency to look for explanations in stable characteristics of individuals, and to underestimate situational variability, may lead us to over-interpret VAMs as indicating a property of the teacher. Related to this is evidence about the stability of estimates from VAMs.

McCaffrey et al. (2009) found year-to-year correlations in value-added measures in the range of 0.2–0.5 for elementary school and 0.3–0.7 for middle school teachers, and show that this is consistent with the findings of previous studies. Interestingly, it is also comparable with the stability of performance estimates for other professions, such as salespersons, university faculty and baseball players.

In discussing school-level value-added estimates, Gorard, Hordosy & Siddiqui (2012) found the correlation between estimates for secondary schools in England in successive years to be between 0.6 and 0.8. They argued that this, combined with the problem of missing data, makes it meaningless to describe a school as 'effective' on the basis of value-added.

## Student ratings

A review of the research on student rating can be found in Burniske & Neibaum (2012). Among their advantages, the authors report previous findings, whereby student ratings are valid, reliable, cost-effective, related to future achievement, valuable for teacher formative feedback and require minimal training. The disadvantages are that results may require different interpretations according to the students' age, and generally the fact that teachers would resist such an assessment if it was solely used for their appraisal.

Student evaluation of teaching is a topic which has been widely explored by higher education research, as it is one of the preferential evaluation methods in the United States and in the United Kingdom, and owes much to the work of Herbert W. Marsh on developing valid and reliable student assessment questionnaires (Marsh, 1982, 2007; Richardson, 2005). Today, most literature agrees that while students' assessment of teaching can be valid and reliable, there needs to be careful use of the plethora of available instruments that can be a tool for formative assessment (Law, 2010; Spooren, Brockx & Mortelmans, 2013).

Much less is known about student ratings in school settings. Mertler (1999) reviews research summarising the benefits of using student observations for measurement purposes (for instance, 'no one is in a better position to critique the clarity of teacher directions than the students for whom the directions are intended', Stiggins & Duke, 1988, cited in Mertler, 1999, pp. 19–20). After testing a purposely-developed feedback questionnaire on nearly 600 secondary students, Mertler (1999) reports that the participating teachers were supportive of the pilot and that student feedback could be a useful measure for teacher formative

assessment. Clearly, the low stakes and the absence of any real follow-up engagement from the teachers should put these results into perspective.

Peterson, Wahlquist & Bone (2000) use data from almost ten thousand students from a school district in the United States. Unlike Merton (1999), the authors rely on a pre-existing evaluation system involving student results as part of a wider appraisal scheme. They find that students 'responded to the range of items with reason, intent, and consistent values' (Peterson, Wahlquist & Bone, 2000, p. 148). Pupil surveys have also been shown to predict achievement in primary education. Drawing from teacher effectiveness research, Kyriakides (2005) uses data from almost 2000 primary school children in Cyprus to show that 'student ratings of teacher behavior are highly correlated with value-added measures of student cognitive and affective outcomes'.

## Principal (headteacher) judgement

Evaluations by principals are typically based on classroom observations, possibly using informal brief drop-in visits. However, principals are also able to draw on considerable background knowledge, both of the individual teacher and of the context in which the evaluation takes place. It may also be that they have access to additional information about the teacher, the effect of which could be either to inform or bias the judgement they make.

Broadly speaking, the research evidence suggests that principal judgements correlate positively with other measures, but the correlations are modest. For example, Jacob and Lefgren (2008) found correlations of around 0.2 between principal ratings of teachers' impact on their students' learning and value-added measures.

## Teacher self-reports

Self-reports include tools such as surveys, teacher logs and interviews. The content of what is reported may vary considerably. The evidence reviewed by Goe et al (2008) about validity and reliability of self-report surveys suggests that they may not currently be trustworthy as a measure of quality. Teacher logs and interviews similarly suffer from low reliability and all these measures have only modest correlations with other measures of effectiveness. Self-report measures of any kind also tend to be influenced by social desirability biases.

## Analysis of classroom artefacts

Analysis of artefacts such as lesson plans, teacher assignments, assessment methods and results, or student work, seems like an obvious way to judge the effectiveness of the teaching. There is some evidence that when raters follow a specific protocol for evaluating these artefacts, the results are reasonably consistent with other measures (Goe et al, 2008).

One such protocol is the Instructional Quality Assessment (IQA). The most work on this has been done by the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) located at the University of California–Los Angeles (Matsumura et al., 2006). Another is the Intellectual Demand

Assignment Protocol (IDAP), developed by Newmann and colleagues of the Consortium on Chicago School Research (Newmann et al., 2001). In both these cases the evidence of validity and reliability comes from studies conducted by the developers. This makes it hard to judge what the performance of the measures might be in regular use in schools.

## Teacher portfolios

Portfolios “are a collection of materials compiled by teachers to exhibit evidence of their teaching practices, school activities, and student progress” (Goe et al, 2008). They may include “teacher lesson plans, schedules, assignments, assessments, student work samples, videos of classroom instruction and interaction, reflective writings, notes from parents, and special awards or recognitions.” An important difference between portfolios and analysis of artefacts, is that the content of the portfolio is selected or created by the individual teacher to show their achievements to best effect. Although it is sometimes claimed that the value of the portfolio is in the reflection that underpins the process, they are also used as a source of evaluation evidence and for certification.

Probably the best known example of the use of teacher portfolios is the National Board certification for its Professional Teaching Standards (NBPTS). NBPTS has been the subject of a substantial amount of research, though the findings are somewhat mixed. Some studies do find a link between portfolio scores and other measures of teaching quality, but others do not. Achieving acceptable inter-rater reliability among markers is also not straightforward (Goe et al, 2008). Despite considerable enthusiasm for this approach in some quarters, the assessment of teacher portfolios as a measure of teaching quality is probably not justified.

## How could this promote better learning?

So far we have reviewed the evidence about what great teaching looks like, and how it can be safely identified. This evidence is important for teachers to understand, but it is in some ways just a preamble to the crucial question of how that understanding can be used to improve students' learning. Before we can do that, we must first clarify some validity issues that arise out of any attempt to 'measure' teaching quality. Then we consider relevant evidence about how feedback about teaching quality can be used most effectively, and how this relates to the broader issue of teachers' professional development.

### Validity Issues

#### Combining evidence from different evaluation approaches

One question we need to address early on is whether we are setting out to produce a single measure of teaching effectiveness. Today, many jurisdictions are using multiple sources of teacher evaluation, but with the intention of combining them into an overall measure (Burniske & Neibaum, 2012; Isoré 2009). A single measure will be required, for example, if we want to rank teachers in order of effectiveness, or to attach explicit consequences to different score ranges. On the other hand, if we want to focus on giving teachers feedback on a range of strengths and weaknesses, such a combined score may be unnecessary and unhelpful.

It may be that part of the reason researchers have not been more successful in achieving congruence across different methods and instruments for assessing effectiveness is that there is not just one kind of effectiveness. It may be, for example, that different teachers with very different sets of skills, knowledge and understanding can achieve similar ends in terms of students' learning. A measurement approach that starts from the assumption that the answer is a weighted sum of all the component parts may miss the subtlety of their interactions. If our investigative method is to feed potential explanatory factors into regression models we will be unlikely to find these kinds of relationships.

There may, for example, be threshold effects, so that once a particular teacher skill reaches an adequate level, further increases do not make much difference; below that level, however, and learning is likely to be diminished. Or there may be interactions, so that two (or more) particular skills can compensate for each other: as long as at least one of them is strong enough, the strength of the other is unimportant.

All of this is speculation, of course: any theory of teaching effectiveness would have to be developed fully and tested. But it may be important to keep an open mind about the kinds of relationships we may find.

#### Focus on student learning outcomes

We have already made clear that our definition of effective teaching is that which leads to enhanced student outcomes. An important corollary is that our criterion measure, against which we should validate all other sources of evidence about

effectiveness (such as from lesson observation, student ratings, etc.) must always be anchored in direct evidence of valued learning outcomes.

We need to stress that this does not mean that we have to privilege current testing regimes and value-added models. Existing measures and models may fall well short of what we need here. However, success needs to be defined not in terms of teacher mastery of new strategies or the demonstration of preferred behaviours, but in terms of the impact that changed practice has on valued outcomes. Because teachers work in such varied contexts, there can be no guarantee that any specific approach to teaching will have the desired outcomes for students.

### Purposes: Fixing versus Firing

A key part of modern thinking about validity is that we need to know the purposes for which a measure is intended to be used before we can evaluate any evidence about whether it is fit for purpose.

James Popham (1988) has characterised two incompatible uses of measures of effectiveness as 'Fixing' (formative assessment, intended to improve practice) and 'Firing' (summative assessment, with consequences attached, e.g. merit pay or termination of employment). He pointed out that either may be fine alone, but together they make a counter-productive 'dysfunctional marriage'.

As Hinchey (2010, p6) explains

> "Assessment to improve practice requires that teachers be open to admitting weaknesses, which can happen only in a relatively non-threatening environment. … Teachers whose work can be improved but who are feeling at risk may understandably be inclined to hide, rather than confront, their problems—precluding valuable formative feedback."

The requirements for a measure to be used for 'fixing' may be very different from those for 'firing'. It will not be helpful to talk about 'validity' in a general sense without being clear about this.

## Approaches to providing feedback

A range of studies suggests that the quality of feedback is a key component of any teacher assessment (Stiggens & Duke (1988), McLaughlin & Pfeifer (1988), Kimball (2002)).

Hattie & Timperley (2007) state that the main purpose of feedback 'is to reduce discrepancies between current understandings and performance and a goal' (ibid., p. 86). Although their review concerns teacher feedback to students, given that learning works in similar ways for adults and young people (Bransford, Brown, & Cocking, 2000) their findings can be adapted for our focus on feedback as a follow-up activity to an observation.

Hattie & Timperley argue that effective feedback answers three questions ('Where am I going?', 'How am I going?' and 'Where to next?') and operates at four levels:

the task ('How well tasks are understood/performed'); process ('the main process needed to understand/perform the task'); self-regulation; and self level ('Personal evaluations and affect […] about the learner').

Timperley et al. (2007) review the characteristics of the teacher 'knowledge-building cycle ' - a feedback loop for teachers - that are associated with improved student outcomes. Their synthesis 'assumes that what goes on in the black box of teacher learning is fundamentally similar to student learning'. Their findings suggest that teacher learning can have a sizeable impact on student outcomes.

They report that in effective interventions feedback was related to evidence and clear goals about developing teacher pedagogical content knowledge and student achievement or conceptual understanding, whilst providing the teacher with the skills to assess student outcomes. Moreover, professional instruction was followed by a range of opportunities to practice and learn.

The observation/feedback routine should be structured explicitly as a continuous professional learning opportunity that actively challenges teacher thinking and practice and enables them to work on improving, for it to be more likely to translate into student outcomes: teacher learning drives student learning. Principals can help by 'developing a vision of how teaching might impact on student outcomes, managing the professional learning environment, promoting a culture of learning within the school, and developing the leadership of others in relation to curriculum or pedagogy.'

**Evidence of impact of feedback to teachers on student learning**

This is some evidence, reviewed by Coe (2002), that the use of feedback information from school performance measures can have positive effects on subsequent school performance. However, as Coe points out, we are limited by the lack of both direct evidence and strong theory:

> Given the complexity of the kinds of feedback that can be given to schools about their performance, the varying contexts of school performance, and the range of ways feedback can be provided, it is extremely difficult to make any kind of generalised predictions about its likely effects.

One specific example of a positive impact of feedback from classroom observation is from Taylor and Tyler (2012). They used Danielson's *Framework for Teaching* to evaluate and feed back to teachers in Cincinnati over a period of seven years. They found a gain in students' performance in math test scores in the years following the intervention, equivalent to an effect size of 0.11. The cost of the observation intervention was estimated at $7,500 per teacher.

## Enhancing teachers' professional learning

Timperley (2008) highlights a number of broad principles from an extensive research review on successful professional learning - and much of this advice can be translated to observation and feedback routines or programmes in general. To be effective, strategies:

- Must focus on and be measured against student outcomes;
- Encourage 'self-regulation' among teachers who need to embrace the experience as independent learners and sustain the techniques;
- Require some input from school leaders;
- Involve, ideally, collaboration with peers;
- Be a genuine challenge.

| | **Summary of advice from Timperley (2008)** |
|---|---|
| **1** | 'Focus on valued student outcomes', whether it is achievement or a deeper student understanding |
| **2** | 'Professional knowledge and skills that do have a positive impact on student outcomes are consistent with evidence-based principles of teaching effectiveness', national associations' recommendations, or with rigorously-debated national policies. |
| **3** | 'To establish a firm foundation for improved student outcomes, teachers must integrate their knowledge about the curriculum, and about how to teach it effectively and how to assess whether students have learned it'. We consider the last point to be especially relevant, as it is the basis for teacher monitoring of students but also self-regulation. |
| **4** | 'To make significant changes to their practice, teachers need multiple opportunities to learn new information and understand its implications for practice.<br><br>Furthermore, they need to encounter these opportunities in environments that offer both **trust** and **challenge**' |
| **5** | Whether the decision of engaging with professional development is voluntary or directed has no bearing on student outcomes. |
| **6** | '[I]f teachers are to change, they need to participate in a professional learning community *that is focused on becoming responsive to students* […]. As an intervention on its own, a collegial community will often end up merely entrenching existing practice and the assumptions on which it is based'. |
| **7** | 'Expertise external to the group of participating teachers is necessary to challenge existing assumptions and develop the kinds of new knowledge and skills associated with positive outcomes for students', and this expertise can come from within or outside the school.<br><br>When it is provided by the principal or other school leaders, these professionals should establish 'a vision of new possibilities […] through everyday activities', lead learning and organise learning opportunities. |
| **8** | 'Sustained improvement in student outcomes requires that teachers have sound theoretical knowledge, evidence-informed inquiry skills, and supportive organizational conditions'. |

One example of the importance of the school context in which professional learning takes place comes from a study by Kraft and Papay (2014). They provide a challenge to the now much quoted claim that teachers typically improve over their first 3-5 years and then plateau (e.g. Rockoff, 2004). Kraft and Papay found on average the same pattern: rapid improvement over the first three years, then much slower growth. However, they also found that teachers working in schools with ‘more supportive’ professional environments (assessed by teacher questionnaires) continued to improve significantly after three years, while teachers in the least supportive schools actually declined in their effectiveness.

## How might we take this forward?

This final section of our review pulls together the implications of the research evidence we have presented and proposes a framework for conceptualising teaching quality. We then make some recommendations for practitioners about how these ideas could be used to promote better teaching.

### Overview of the evidence

#### Evidence about effective pedagogy

In Section 2 (p9) we identified a selection of teaching approaches, skills and knowledge that have been shown to be related to enhanced student outcomes. The evidence here is often weak or equivocal, and it is easy to select from it to make claims that fit preconceptions. The effective practices themselves are often quite loosely described, leaving room for interpretation about whether what one has observed is in fact an example of it. Partly for this reason, we also provided a list of ineffective practices: teaching approaches that seem to be popularly endorsed by at least some teachers, but whose use is not supported by research (p22).

How teaching leads to learning is undoubtedly very complex. It may be that teaching will always be more of an art than a science, and that attempts to reduce it to a set of component parts will always fail. If that is the case then it is simply a free-for-all: no advice about how to teach can claim a basis in evidence. However, the fact that there are some practices that have been found to be implementable in real classrooms, and that implementing them has led to improvements in learning, gives us something to work with. Much of this work is under-theorised and difficult to make sense of. However, the Dynamic Model of Creemers and Kyriakides (2006) provides a theory that is well specified and has withstood some credible attempts to test it. For now at least, it is the best theory of effective pedagogy we have.

#### Evidence about methods of evaluating teaching quality

The rise of accountability pressures in many parts of the world have led to a big growth in the desire to evaluate the quality of teaching. A number of methods have been widely used and evaluated in research studies.

Value-added models are highly dependent on the availability of high-quality outcome measures. Their results can be quite sensitive to some essentially arbitrary choices about which variables to include and how to fit the models. Estimates of effectiveness for individual teachers are only moderately stable from year to year and class to class. However, it does seem that at least part of what is captured by value-added estimates does reflect the genuine impact of a teacher on students' learning.

Classroom observation seems to have face validity as an evaluation method, but the evidence shows that the agreement between different observers who see the same lesson is not high; neither is agreement between estimates of teaching quality from lesson observation and from other methods. Levels of reliability that

are acceptable for low-stakes purposes can be achieved by the use of high-quality observation protocols, use of observers who have been specifically trained – with ongoing quality assurance – in using those protocols, and pooling the results of observations by multiple observers of multiple lessons (Strong et al, 2011, Mihaly et al, 2013).

There is some evidence that principals' judgements about teacher quality have positive but modest correlations with other evidence. Inferring the quality of teaching and learning from looking at artefacts such as student work, marking or lesson plans, or from assessing teacher portfolios, is not currently supported by research as valid.

### Evidence about developmental use of evaluation

The assessment of teaching quality need not necessarily have summative evaluation as its aim. Indeed, our focus in this review is primarily on formative uses of assessment. In designing systems to support such uses, we need to take account of the characteristics of feedback that are most likely to lead to positive effects and of the environment in which the feedback is given and received.

Specifically, feedback should relate performance to clear, specific and challenging goals for the recipient. It should direct attention to the learning rather than to the person or to comparisons with others. Feedback is most likely to lead to action when it is mediated by a mentor in an environment of trust and support. Sustained professional learning is most likely to result when the focus is kept clearly on improving student outcomes, when there are repeated and sustained opportunities to embed any learning in practice, when the promotion of new thinking about teaching takes account of existing ideas, and when an environment of professional learning and support is promoted by the school's leadership.

## A general framework for teaching quality

A number of frameworks for conceptualising the elements of effective teaching have been presented. Broadly speaking they include the following components:

*2. (Pedagogical) content knowledge*

The evidence to support the inclusion of content knowledge in a model of teaching effectiveness is strong, at least in curriculum areas such as maths, literacy and science. Different forms of content knowledge are required. As well as a strong, connected understanding of the material being taught, teachers must also understand the ways students think about the content, be able to evaluate the thinking behind non-standard methods, and identify typical misconceptions students have.

*5. Quality of instruction*

Quality of instruction is at the heart of all frameworks of teaching effectiveness. Key elements such as effective questioning and use of assessment are found in all of them. Specific practices like the need to review previous learning, provide models for the kinds of responses students are required to produce, provide

adequate time for practice to embed skills securely and scaffold new learning are also elements of high quality instruction.

*4. Classroom climate / relationships / expectations*

Again, the empirically based frameworks all include something on classroom climate, though this heading may cover a range of aspects of teaching. Some (e.g. CLASS) emphasise the quality of relationships and interactions between teachers and students. Also under this heading may come teacher expectations: the need to create a classroom environment that is constantly demanding more and never satisfied, but still affirming to students' self-worth and not undermining their feelings of self-efficacy. Promotion of different kinds of motivational goals may also fit here, as may the different attributions teachers make and encourage for success and failure (e.g. fixed versus growth mindset, attributions to effort and strategy rather than ability or luck). Related to this is the valuing and promotion of resilience to failure (grit).

*3. Behaviour / control / classroom management*

All the empirically based frameworks include some element of classroom management. A teacher's abilities to make efficient use of lesson time, to coordinate classroom resources and space, and to manage students' behaviour with clear rules that are consistently enforced, are all relevant to maximising the learning that can take place. These factors are mostly not directly related to learning; they are necessary hygiene factors to allow learning, rather than direct components of it.

*1. Beliefs (theory) about subject, learning & teaching*

The idea that it matters why teachers adopt particular practices, the purposes they aim to achieve, their theories about what learning is and how it happens and their conceptual models of the nature and role of teaching in the learning process all seem to be important. Although the evidence to support this claim is not unequivocal, it seems strong enough to include it at this stage.

*6. Wider professional elements: collegiality, development, relationships*

It seems appropriate to include a final heading that captures some broader aspects of professional behaviour. Danielson's Framework for Teaching includes elements such as reflecting on and developing professional practice, supporting colleagues, and liaising and communicating with stakeholders such as parents. There may not be direct evidence linking these practices to enhanced student outcomes, but if we want to capture a broad definition of effective teaching, they should probably be included.

## Best bets to try out and evaluate

Any recommendations we make here are tentative and very likely to be modified. Crucially as well, we must build in robust evaluation into any changes we make; any recommendations are only hypotheses about what might help. Nevertheless, it is important at least to try to capture some suggestions about how we can take these ideas forward to enhance learning. Some actions will be easier than others,

so we have divided them into quick wins and longer term changes. First, though, we outline some general requirements for system improvement.

### General requirements

There are a few general requirements that follow from the previous arguments. The first is that a worthwhile system for monitoring and formative evaluation of teaching quality must have at its heart a set of high-quality assessments of student learning. Building in assessment ensures that we keep the focus on student outcomes. If the assessments are of high-quality that ensures that they will capture the learning outcomes that we value and want to incentivise. Ultimately, for a judgement about whether teaching is effective to be seen as trustworthy, it must be checked against the progress being made by learners. However good our proxy measures become, there is no substitute for this.

A second requirement is that a formative teacher evaluation system must incorporate multiple measures, from multiple sources, using multiple methods. Users must triangulate multiple sources of evidence, treating each with appropriate caution, critically testing any inferences against independent verification. The more sources of evidence we have, the better our judgements can be.

A third requirement, related to these two, is the need for a high level of assessment and data skills among school leaders. The ability to identify and source 'high-quality' assessments, to integrate multiple sources of information, applying appropriate weight and caution to each, and to interpret the various measures validly, is a non-trivial demand.

A fourth and final requirement is the need to balance challenge and acceptance. If the gap between research-based 'effective practices' or data from performance evaluation and existing perceptions is too big the former are likely to be rejected. On the other hand, if the requirements are perceived to be similar to current practice, nothing will change. The latter would be an example of the 'we think we are doing that' problem: teachers take on superficial aspects of a new approach, or interpret current practice as aligned with it, and an opportunity for improvement is lost.

### Quick wins

A number of specific recommendations should be possible for teachers to implement quickly and without great cost:

1. Spread awareness of research on effective pedagogy.
   The evidence that has been presented in Section 0 about effective teaching approaches may not be universally known by teachers. We should encourage all teachers to engage with these ideas, to challenge their own thinking and that of their colleagues about what is effective, and to understand the kind of evidence that supports the claims.

2. Use the best assessments available.
   Ultimately, the definition of effective teaching is that which results in the

best possible student outcomes. There is currently no guaranteed recipe for achieving this: no specifiable combination of teacher characteristics, skills and behaviours consistently predicts how much students will learn. It follows that the best feedback to guide the pursuit of effectiveness is to focus on student progress, and that requires high-quality assessment of learning.

3. Use lesson observation, student ratings, artefacts and principal judgement cautiously.
   All these methods have potential value, but all have their problems. If they are done well, using the best available protocols, with awareness of how they can be biased or inaccurate, and with due caution about what inferences they can and cannot support, then they should be useful tools.

4. Triangulate.
   A key to suitably cautious and critical use of the different methods is to triangulate them against each other. A single source of evidence may be suggestive, but when it is confirmed by another independent source it starts to become credible. Having more data can sometimes make people feel overwhelmed and indecisive, but for anyone who truly understands the limitations of a single source, being restricted to that would feel hopelessly exposed.

5. Follow the advice from Timperley (2008) about promoting professional learning.
   Sustained professional learning is most likely to result when the focus is kept clearly on improving student outcomes, when there are repeated and sustained opportunities to embed any learning in practice, when the promotion of new thinking about teaching takes account of existing ideas, and when an environment of professional learning and support is promoted by the school's leadership.

**Longer term (harder)**

In addition to these quick wins, there are other recommendations that may be harder, take longer or cost more to implement. There are broadly two kinds of approaches here: one focuses on developing the measures we need to evaluate effectiveness robustly, the other on developing the support systems that promote the use of feedback for improvement.

**Multiple, multi-dimensional measures**

If the measures we need do not exist, it may be necessary to create them. If they do exist, but are not yet ideal for our purposes, it may be necessary to develop them further. If they already exist in a suitable format, then we still need to validate them against our criteria for developmental consequences: does using them as part of a formative evaluation process for teachers lead to improved student outcomes?

*Create better assessments*

In order to judge the effectiveness of their teaching, teachers need to have access to assessments that reflect the learning they are trying to promote, that are calibrated to allow judgements about expected rates of progress, that cover the full range of curriculum areas and levels, and that are cheap and easy to administer on a frequent basis. Although generally of high psychometric quality, available standardised tests do not routinely meet all these requirements.

It may be that system of crowd-sourced assessments, peer-reviewed by teachers, calibrated and quality assured using psychometric models, and using a range of item formats, could meet this need.

*Lesson observation tools*

A number of protocols exist for lesson observation, and it may be that the best of them provide an optimal way forward. However, it may also be that their requirements for training are prohibitively onerous or expensive, or that alternatives could be developed that better meet the needs of a developmental focus, that are led and owned by the profession, and that make best use of online communities for video sharing, peer ratings and maximising learning for both observed and observer.

One example would be a simple tool for measuring students' time on task in lessons. Brophy and Good (1986, p360) identify the relationship between 'academic engaged time' and student achievement as one of the 'most consistently replicated findings' in the literature. Giving a teacher this relatively objective measure and allowing them to track its trajectory over time and with different classes, perhaps contextualised against the values that other teachers achieve with similar students, could be an effective way to increase the percentage of time spent engaged in lessons and hence to improve learning.

*Student ratings*

Again, these instruments exist, so this could actually be quite a quick win. Collecting student ratings should be a cheap and easy source of high-quality feedback about teaching behaviours from multiple observers who can draw on experience of multiple lessons. Although there is evidence of using student ratings to enhance learning outcomes in higher education, their use in schools does not appear to have been evaluated yet.

**School-based support systems**

Creating systems of support within schools that would allow teachers to respond positively to the challenge of improving their effectiveness is an important task. There are many advantages to a school-led system here: it keeps the ownership within the profession and makes the whole process more straightforward to manage. One danger is that without some external expertise the learning may be limited to what is already available in-house (Antoniou & Kyriakides, 2011). It may also be hard to create high challenge in a peer-to-peer system. Part of the reason for generating objective measures of a range of aspects of teaching effectiveness is that they provide an external check against which to compare.

*Mentoring*

There are many existing models of school-based professional mentoring, so it should not be difficult to select a small number of promising ones for this purpose and evaluate their impact. Key design issues include creating mentoring relationships characterised by trust and feeling supported, while being sufficiently challenging to provoke change. The difficulties of sustaining real change over a long period should also be addressed in the design.

*Lesson Study*

Another possible route would be to use a Lesson Study approach. Originally from Japan, it was imported in the United States and the United Kingdom and involves groups of teachers collaboratively planning, teaching, observing and analyzing learning and teaching in 'research lessons'. (Dudley, 2014, p. 1)

In the United States, Lesson Study was found to be one of the two interventions, out of the many hundreds systematically reviewed, to have statistically significant positive effects on the pupils' fraction knowledge in grades 2, 3 and 5 (Gersten et al., 2014). Cajkler et al. (2014) argue that Lesson Study provides four benefits: 'Greater teacher collaboration'; 'sharper focus among teachers on students' learning'; 'development of teacher knowledge, practice and professionalism'; and 'improved quality of classroom teaching and pupil learning outcomes.' (ibid., p. 3).

Dudley (2014) suggests that the reasons why Lesson Study works are that it is a gradual process that places specific learners' needs as a focus for development. It involves an element of collaborative enquiry or experiment between teachers who are trying to solve a problem and that takes place 'in the context of a supportive teaching and learning community'. There is also input from an external expertise. In all studies finding positive effects from the implementation of Lesson Study, a considerable role was played by an agent outside the teacher group that could provide feedback and challenge their views.

As with other feedback programmes Lesson Study faces a number of challenges. Saito et al, (2008) report varied opinions among the faculty members with regard to how to observe lessons. Teacher groups 'also differ[ed] in terms of the types of discussions during reflection', with some focusing more on the teaching process and others on student behaviours. Often senior managers or external experts were not involved. Some argue that experiments with Lesson Study may become a practice of 'the blind leading the blind'. This is not a negligible point, and it is one of the main recent critiques to those professional development approaches emphasising practitioners' reflection without providing them with a solid theoretical framework of reference against which to assess them (Antoniou & Kyriakides, 2011).

## References


Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system. *Educational Researcher*, 37(2), 65–75.

Askew, M., Brown, M., Rhodes, V., Wiliam, D., & Johnson, D. (1997). Effective Teachers of Numeracy: Report of a study carried out for the Teacher Training Agency. London: King's College, University of London

Ball D.L and Hill H.C. (2009). Measuring teacher quality in practice In D.H. Gitomer (Ed.), *Measurement issues and assessment for teaching quality*. Washington, DC: Sage Publications.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M.A. Gernsbacher, et al (Ed) *Psychology and the real world: Essays illustrating fundamental contributions to society*, New York: Worth Publishers (p56-64). [Avaliable at http://bjorklab.psych.ucla.edu/pubs/EBjork_RBjork_2011.pdf ]

Boaler, J. (2008). When politics took the place of inquiry: A response to the National Mathematics Advisory Panel's review of instructional practices. *Educational Researcher*, 37(9), 588-594.

Bouchamma, Y. 2005. Evaluating teaching personnel. Which model of supervision do Canadian teachers prefer? *Journal of Personnel Evaluation in Education, 18*(4), pp. 289–308.

Bramschreiber, T. (2012). Taking Peer Feedback to Heart. *Educational Leadership, 70*(3), retrieved from http://www.ascd.org/publications/educational-leadership/nov12/vol70/num03/Taking-Peer-Feedback-to-Heart.aspx

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience, and school*: Expanded edition. Washington, DC: National Academy Press. http://www.colorado.edu/MCDB/LearningBiology/readings/How-people-learn.pdf

Brophy, J., Good, T. (1986). Teacher behavior and student achievement. In M. Wittrock (Ed.), *Handbook of Research on Teaching* (pp. 328 – 375). New York: Macmillian.

Brown, P. C., Roediger III, H. L., & McDaniel, M. A. (2014). *Make it Stick: The Science of Successful Learning*. Harvard University Press.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood (NBER Working Paper No. w17699). National Bureau of Economic Research.

Coe, R. (2002) 'Evidence on the Role and Impact of Performance Feedback in Schools' in A.J.Visscher and R. Coe (eds.) *School Improvement Through Performance Feedback*. Rotterdam: Swets and Zeitlinger

Creemers, B. P. M., & Kyriakides, L. (2006). Critical analysis of the current approaches to modelling educational effectiveness: The importance of establishing a dynamic model. *School Effectiveness and School Improvement*, 17, 347–366. http://www.rug.nl/staff/b.p.m.creemers/testing_the_dynamic_model_of_educational_effectiveness.pdf

Creemers, B. P., & Kyriakides, L. (2011). *Improving Quality in Education: Dynamic Approaches to School Improvement*. Routledge, Taylor & Francis Group.

Danielson, C. (2007). Enhancing professional practice: A framework for teaching. ASCD.

Danielson, C., & McGreal, T. L. (2000). Teacher evaluation to enhance professional practice. ASCD.

Stipek, D. (2010) *How Do Teachers' Expectations Affect Student Learning* http://www.education.com/reference/article/teachers-expectations-affect-learning/

Dumay, X., Coe, R., and Anumendem, D. (2013) 'Stability over time of different methods of estimating school performance'. *School Effectiveness and School Improvement*, vol 25, no 1, pp 65-82.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4-58.

Dweck, C.S. (1999) 'Caution – Praise Can Be Dangerous' *American Educator*, Spring 1999, p4-9. https://www.aft.org/pdfs/americaneducator/spring1999/PraiseSpring99.pdf

Faubert, V. (2009). School evaluation: current practices in OECD countries and a literature review. *OECD Education Working Papers*, No. 42, OECD Publishing. doi:10.1787/218816547156.

Geake, J. G. (2008) Neuromythologies in education. Educational Research 50, 123–133.

Gersten, R., Taylor, M. J., Keys, T. D., Rolfhus, E., & Newman-Gonchar, R. (2014). *Summary of research on the effectiveness of math professional development approaches*. (REL 2014–010). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast.

Glickman, C.D. 1990. *Supervision of instruction: A developmental approach*. 2nd ed. Needham Heights, MA: Allyn and Bacon.

Goe, L., Bell, C., & Little, O. (2008). Approaches to Evaluating Teacher Effectiveness: A Research Synthesis. National Comprehensive Center for Teacher Quality. http://files.eric.ed.gov/fulltext/ED521228.pdf

Goldhammer, R. 1969. Clinical supervision. Special methods for the supervision of teachers. New York, NY: Holt, Reinhart, and Winston

Goldstein, J. (2007). Easy to dance to: Solving the problems of teacher evaluation with peer assistance and review.*American Journal of Education, 113*(3), 479-508.

Gorard, S., Hordosy, R., & Siddiqui, N. (2012). How Unstable are 'School Effects' Assessed by a Value-Added Technique?. *International Education Studies*, 6(1), p1.

Gorard, S., See, B. H., & Davies, P. (2012). The impact of attitudes and aspirations on educational attainment and participation. *York: Joseph Rowntree Foundation. Available at:* http://www.jrf.org.uk/sites/files/jrf/education-young-people-parents-full.pdf.

Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The Test Matters: The Relationship Between Classroom Observation Scores and Teacher Value Added on Multiple Types of Assessment. *Educational Researcher*, *43*(6). 293–303.

Hamre, B.K., Goffin, S.G. & Kraft-Sayre, M. (2009) Classroom Assessment Scoring System Implementation Guide: Measuring and Improving Classroom Interactions in Early Classroom Settings. http://www.teachstone.org/wp-content/uploads/2010/06/CLASSImplementationGuide.pdf

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research*, 77(1), 81-112.

Higgins, S., & Moseley, D. (2001). Teachers' thinking about information and communications technology and learning: Beliefs and outcomes. *Teacher Development*, 5(2), 191-210.

Higgins, S., Katsipataki, M., Kokotsaki, D., Coleman, R., Major, L.E., & Coe, R. (2013). The Sutton Trust-Education Endowment Foundation Teaching and Learning Toolkit. London: Education Endowment Foundation. [Available at http://www.educationendowmentfoundation.org.uk/toolkit ]

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, 48(3), 794-831.

Hill, H. C., Rowan, B., and Ball, D. L. (2005). Effects of Teachers' Mathematical Knowledge for Teaching on Student Achievement. *American Educational Research Journal*, 42(2), 371-406.

Hinchey, P. H. (2010). Getting Teacher Assessment Right: What Policymakers Can Learn from Research. National Education Policy Center. http://nepc.colorado.edu/publication/getting-teacher-assessment-right

Hobson, A. J., & McIntyre, J. (2013). Teacher fabrication as an impediment to professional learning and development: the external mentor antidote. *Oxford Review of Education*, *39*(3), 345-365.

Howard-Jones, P.A. (2014) Neuroscience and education: myths and messages. *Nature Reviews Neuroscience*. Advanced Online Publication, published online 15 October 2014; doi:10.1038/nrn3817. http://www.nature.com/nrn/journal/vaop/ncurrent/pdf/nrn3817.pdf

Husbands, C., & Pearce, J. (2012). What is great pedagogy: A literature review for the National College of School Leadership. National College for School Leadership.

Isoré, M. (2009). Teacher evaluation: current practices in OECD countries and a literature review. OECD Education Working Papers, No. 23 http://hdl.handle.net/123456789/2541

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*, *26*(1), 101-136.

James, M. and Pollard, A. (2011) TLRP's ten principles for effective pedagogy: rationale, development, evidence, argument and impact. *Research Papers in Education*, 26 (3). pp.

Johnson S. M., Fiarman, S. E, Munger, M. S., Papay, J. P., & Qazilbash, E. K. (2009). *A user's guide to Peer Assistance and Review*. Retrieved from http://web.archive.org/web/20100812220001/http://www.gse.harvard.edu/~ngt/par/resources/users_guide_to_par.pdf

Kahlenberg, R. D. (2007). Peer Assistance and Review. Retrieved from http://www.aft.org/newspubs/periodicals/ae/fall2007/kahlenbergsb1.cfm

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have we identified effective teachers? Validating measures of effective teaching using random assignment (MET Project Research Paper). Seattle, WA: Bill & Melinda Gates Foundation.

Kennedy, M. M. (2010). Attribution error and the quest for teacher quality. *Educational Researcher*, 39(8), 591-598.

Kirschner, P. A., Sweller, J., and Clark, R. E. (2006) Why minimal guidance during instruction does not work: an analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. Educational Psychologist 41 (2) 75-86.

Ko J., Sammons P. and Bakkum, L. (2013) Effective Teaching: a review of research and evidence. CfBT Education Trust. http://cdn.cfbt.com/~/media/cfbtcorporate/files/research/2013/r-effective-teaching-2013.pdf

Kohut, G. F., Burnap, C., & Yon, M. G. (2007). Peer observation of teaching: Perceptions of the observer and the observed. *College Teaching, 55*(1), 19-25.

Kraft, M. A., & Papay, J. P. (2014). Can Professional Environments in Schools Promote Teacher Development? Explaining Heterogeneity in Returns to Teaching Experience. Educational Evaluation and Policy Analysis, 0162373713519496.

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee value added assessment system. Educational Evaluation and Policy Analysis, 25, 287–298.

Kyriakides L., Creemers B.P.M. & Antoniou P. (2009) Teacher behaviour and student outcomes: Suggestions for research on teacher training and professional development. *Teaching and Teacher Education*, 25, 12-23.

Kyriakides, L., Christoforou, C., & Charalambous, C. Y. (2013). What matters for student learning outcomes: A meta-analysis of studies exploring factors of effective teaching. Teaching and Teacher Education, 36, 143–152.

Kyriakides, L., Creemers, B., Antoniou, P., & Demetriou, D. (2010). A synthesis of studies searching for school factors: Implications for theory and research. British Educational Research Journal, 36, 807–830.

Little, O., Goe, L., & Bell, C. (2009). A Practical Guide to Evaluating Teacher Effectiveness. National Comprehensive Center for Teacher Quality.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, 44(1), 47–67.

Lord, B., Cress, K., & Miller, B. (2008). Teacher leadership in support of large-scale mathematics and science education reform. In M.M. Mangin & S.R. Stoelinga (Eds.), *Effective teacher leadership: Using research to inform and reform.* New York: Teachers College Press.

Marsh, H. W. (1982) SEEQ: a reliable, valid and useful instrument for collecting students' evaluations of university teaching, *British Journal of Educational Psychology, 52*, 77–95.

Marsh, H.W. (2007), Students' evaluations of university teaching: dimensionality, reliability, validity, potential biases and usefulness. In Perry, R.P. and Smart, J.C. (Eds), *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective*. Berlin: Springer, pp. 319-83.

Marshall, J. C., Horton, B., & White, C. (2009). EQUIPping teachers: A protocol to guide and improve inquiry-based instruction. *The Science Teacher, 76*(4), 46–53.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). Evaluating Value-Added Models for Teacher Accountability. Monograph. RAND Corporation. PO Box 2138, Santa Monica, CA 90407-2138. http://www.rand.org/pubs/monographs/MG158.html

McCaffrey, D., Sass, T., Lockwood, J., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates, *Education Finance and Policy*, 4(4), 572-606. http://dx.doi.org/10.1162/edfp.2009.4.4.572

Mihaly, K., McCaffrey D. F., Staiger D. O., and Lockwood J. R. (2013) A Composite Estimator of Effective Teaching: Report for the Measures of Effective Teaching Project, Bill and Melinda Gates Foundation. Available at http://www.metproject.org/downloads/MET_Composite_Estimator_of_Effective_Teaching_Research_Paper.pdf

Muijs, D, Kyriakides L, van der Werf G, Creemers B, Timperley H & Earl L (2014) State of the art – teacher effectiveness and professional learning, *School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice*, 25:2, 231-256

National Mathematics Advisory Panel (2008) Foundations for Success: The Final Report of the National Mathematics Advisory Panel. March 2008, U.S. Department of Education, Washington, DC. http://www2.ed.gov/about/bdscomm/list/mathpanel/report/final-report.pdf

Newton, X. A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-Added Modeling of Teacher Effectiveness: An Exploration of Stability across Models and Contexts. *Education PolicyAnalysis Archives*, 18(23), n23.

Papay, J. P. (2011). Different Tests, Different Answers: The Stability of Teacher Value-Added Estimates Across Outcome Measures. *American Educational Research Journal*, 48(1), 163-193.

Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2008). Learning styles: concepts and evidence. *Psychological Science in the Public Interest*, 9(3), 105-119. http://www.psychologicalscience.org/journals/pspi/PSPI_9_3.pdf

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom assessment scoring system*. Baltimore: Paul H. Brookes.

Polikoff, M. S. (2014). Does the Test Matter? Evaluating teachers when tests differ in their sensitivity to instruction. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.). Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project (pp. 278-302). San Francisco, CA: Jossey-Bass.

Popham, W.J. (1998). The dysfunctional marriage of formative and summative teacher evaluation. Journal of Personnel Evaluation in Education, 1: 269-273.

Popham, WJ and Ryan JM (2012) Determining a high-stakes test's instructional sensitivity. Paper presented at the annual meeting of the National Council on Educational Measurement, April 12-16, Vancouver, B.C., Canada.

Raudenbush, S.W. (2004), 'What Are Value-added Models Estimating and What Does This Imply for Statistical Practice?', *Journal of Educational and Behavioral Statistics* 29(1):121–129.

Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. Education, 4(4), 492-519.

Riener, C., & Willingham, D. (2010). The myth of learning styles. Change: The magazine of higher learning, 42(5), 32-35. http://new.peoplepeople.org/wp-content/uploads/2012/07/The-Myth-of-Learning-Styles.pdf

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. The American Economic Review, 94(2), 247-252.

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can you recognize an effective teacher when you recruit one? *Education*, 6(1), 43-74.

Rosenshine, B. (2010) Principles of Instruction. International Academy of Education, UNESCO. Geneva: International Bureau of Education. http://www.ibe.unesco.org/fileadmin/user_upload/Publications/Educational_Practices/EdPractices_21.pdf

Rosenshine, B. (2012) Principles of Instruction: Research based principles that all teachers should know. *American Educator*, Spring 2012. http://www.aft.org/pdfs/americaneducator/spring2012/Rosenshine.pdf

Sadler, P. M., Sonnert, G., Coyle, H. P., Cook-Smith, N., & Miller, J. L. (2013). The influence of teachers' knowledge on student learning in middle school physical science classrooms. *American Educational Research Journal*, 50(5), 1020-1049.

Shulman, L. S. (1988). A Union of Insufficiencies: Strategies for Teacher Assessment in a Period of Educational Reform. *Educational Leadership*, *46*(3), 36-41.

Shulman, L. (2009). Assessment of teaching or assessment for teaching? In D.H. Gitomer (Ed.), Measurement issues and assessment for teaching quality. Washington, DC: Sage Publications.

Strong, M., Gargani, J., & Hacifazlioglu, O. (2011). Do we know a successful teacher when we see one? Experiments in the identification of effective teachers. Journal of Teacher Education, 62(4), 367–382. [Abstract at: http://jte.sagepub.com/content/62/4/367.abstract]

Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *The American Economic Review*, *102*(7), 3628-3651.

Timperley, H., Wilson, A., Barrar, H. & Fung, I. (2007) Teacher professional learning and development: Best evidence synthesis iteration. Wellington, New Zealand: Ministry of Education. http://www.educationcounts.govt.nz/publications/series/2515/15341

Timperley, H. (2008). Teacher professional learning and development. International Academy of Education, International Bureau of Education, UNESCO. http://www.ibe.unesco.org/fileadmin/user_upload/Publications/Educational_Practices/EdPractices_18.pdf

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research*, 73(1), 89–122.

Willingham, D. T. (2008). What Will Improve a Student's Memory?. *American Educator*, *32*(4), 17-25.

# Appendix

## A: Original research questions

In more detail, the study set out to review the evidence from existing research to address the following questions:

1. What are the elements of teaching effectiveness and classroom/teaching quality?

    a. What does the educational effectiveness[7] literature say about the factors/characteristics/behaviours of teachers/teaching that are associated with high student attainment/progress?

    b. What is the evidence from intervention studies (eg RCTs) about the classroom strategies that produce increased attainment/progress?

    c. What evidence from psychology (eg on learning, memory, neuropsychology) indicates pedagogical practices that are most likely to lead to deep understanding and retained knowledge?

2. What frameworks/protocols exist for measuring classroom/teaching quality (including use of video and student surveys)?

    a. What frameworks/protocols have been used in research studies? What evidence is there of how effectively these frameworks capture real quality? To what extent are they aligned with the evidence reviewed in 1?

    b. What frameworks/protocols have been used in schools (by practitioners) around the world for measuring teacher effectiveness/quality? What evidence is there of how effective and reliable these frameworks are? To what extent are they aligned with the evidence reviewed in 1?

    c. What requirements for training, accreditation and quality assurance do these frameworks have?

3. In what ways have these frameworks been used in practice to improve practice?

    a. What kinds of outputs/reporting have been developed for these frameworks/protocols?

    b. What models of observer/observed have been tried (eg peer-to-peer, self-evaluation, principal/line manager, external evaluator), and how have professionals collaborated on this?

[7] Eg school effectiveness research, and the 'process-product' literature that looks for correlations between school or classroom processes and outcomes

c. What models of feedback/dialogue and improvement mechanisms exist (eg appraisal/evaluation, information for self-evaluation, support and ‘consultation’ in interpreting and responding to feedback, goal-setting + feedback, etc)?

4. What evidence is there of the impact of any of these approaches on student outcomes?

    a. What high-quality (eg RCT) evaluations exist of interventions based on feedback of classroom quality evaluation?

    b. What claims exist, based on case-studies or other less-rigorous designs? What relevant work is currently underway?