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.