What is wrong with this image?

A plethora of teacher observation schedules (Danielson, Marzano, AITSL) are all about us, and they appear to be the next “golden bullet” for policy makers. I want to argue in this session that most of these are premised on the wrong “focus of analysis”. They are mistakenly based on observing the teacher ….

When they should, instead, focus on the impact of the teachers on student learning – a major difference in focus.

Let me start to build the case for this critical change in focus by summarising some of the research from the meta-analyses in the Visible Learning work about the impact of teachers on learning.

So far I have synthesised over 1100 meta-analyses about various influences on student achievement, based on about 65,000 studies and including about ¼ billion students. Yes, the focus is on achievement outcomes and there are many more outcomes (affective, investment in learning, values, health, social).

The major finding – is that almost everything works if you set the bar at asking about What works to enhance student achievement! Thus, it should not be surprising that most evaluation systems of
teachers can find evidence to support what they do! No wonder every teacher believes that their context is unique, no wonder every teacher (rightly) claims that they can find evidence to show that their “unique” teaching works, and no wonder every academic can find support for their own pet theory or influence about what schools should do.

We have set the bar so low (d > 0) that we have a profession based on “everything works”. Instead, we should raise the bar... to at least the average of all effects (d > 0.40) and ask what are the attributes of teachers who consistently attain greater than average effects - compared to those who attain below average effects – and this was the purpose of the four books on this topic).

For today, the message is that we now have an explanation as to why we can always find evidence of impact from any teacher observation (or any evaluative system).

This is NOT to say that there are not high and low probability teaching methods.

But as Scriven has argued we should not confuse correlates with causes. For example, a teacher could use a low probability method and have excellent results or use a high probability method and be poor at implementing it. It is the impact not the mere presence of excellent “teaching methods” that makes the difference.
Observations based on watching teachers teach are thus fraught with the problems of confusing teaching method with learning impact. Plus – there is a plethora of problems with the current observation methods:

1. Too molecular as to what teacher do. Take the well-known Danielson schedule – so much of what is observed is teacher related and their presence may have no impact on students (although to be fair there is a smaller proportion that are teacher impact focused).

### Teacher focused
- The teacher cites intra- and interdisciplinary content relationships
- The teacher plans demonstrate awareness of possible student misconceptions and how they can be addressed
- The teacher's plans reflect recent developments in content-related pedagogy
- The teacher seeks out information from a students about their cultural heritages
- The teacher maintains a system of updated student records and incorporates medical and/or learning needs into lesson plans
- The teacher has ongoing relationships with colleges and universities that support student learning
- The teacher manages a log of resources for student reference
- Lesson plans differentiate for individual student needs
- Assessments provide opportunities for student choice
- Teacher designed assessments are authentic, with real-word applications as appropriate
- There is total alignment between the learning activities and the physical environment

### Teacher impact focused
- Students are actively involved collecting information from formative assessments and provide input
- Students assist their classmates in understanding the content
- Students take initiative in improving the quality of their work
- Students themselves ensure that transitions another routine are accomplished smoothly
- Students suggest other strategies they might use in approaching a challenge or analysis
- Students initiate higher-order questions
- Virtually all students are engaged in discussion

2. Depends on nature of incoming students (Whitehurst, Chingos, & Lindquist, 2014). Teachers who work with high achieving cohorts are more likely to be rated in the top 20% of all teachers compared to those who work with the lowest cohort of students.
3. Reliability – There needs at be least 3 lessons coded by at least 2 raters to achieve minimal acceptable levels of reliability (Hill, Charalambous & Kraft, 2012). It is noted, however, that observation scores have higher year-to-year reliability than value-added scores (Whitehurst, Chingos & Lindquist, 2014; r=.65 vs. r = .38). And the year-to-year value added in teacher impact still higher than year-to-year baseball averages.

4. Halo effects – Impressions linger such that after observing what is rated as top teaching then the next observations (even of different teachers) tend to be rated higher; and similarly if observing low quality teaching on the next observations.

5. Who observes – For example, administrators’ rate teachers about 10% higher than other administrators and about 20% higher than peers independent raters (Ho & Kane, 2014).

6. Rater content expertise. Raters with expertise in the content are more likely to include observation on the content matter (pedagogical content matter). Most raters do not comment on the content (Goldsmith & Reed, 2014).

7. Voice and clothes among the highest biases.

8. Criteria of success. The majority (80-90%+) of classroom activities can be completed by students using surface not deep knowledge (Fullan, 2014), 90%+ of items in US state assessments can be completed using surface rather than deep knowing – is this the same for observation (i.e., we are privileging teaching of surface not deep). Other outcomes that may relate to differential teaching methods include: reinvestment in learning improving, increasing motivation to learn, mastery, attain depth etc.
9. Restriction of range – most observers are reluctant to use the full range. For example Ho and Kane (2013) found that only 5% were rated as unsatisfactory and 2% as highest (advanced) – the rest were rated in the two middle categories (basic and proficient).

10. The weighting problem. There can many dimensions but how they are weighted to either form an overall impression or to relate to outcomes is critical. For example, Kane, Taylor, Tyler, Wooten (2010) based on large sample of teacher observations (N=2071 teachers) that ‘Creating an environment for student learning’ and ‘teaching for student learning’ is much more critical than ‘planning and preparation and professionalism’ (from the Danielson). Similarly, they found that teachers who are equally adept at “routinized content and standard focused teaching” then the teacher who adds pedagogy that utilizes “questioning and discussion” practices will generate higher reading achievement.

11. Reporting back can be poor. There is a limited research base on optimal reporting such that the interpretations by teachers are both correct and have consequences (Hattie, 2009. 2013). Most feedback can only be generic as this is the nature of most observation instruments, (Hill & Grossman, 2014), and teachers (as humans) are graziers are more likely to interpreted positive attributes about them and negative about conditions which affect their teaching.

12. Can you observe teacher when the teacher is silent and the children are learning? The Visible Learning research shows the critical importance of the decision making processes that teachers use to make decisions about teaching. The VL work is based on decisions about Diagnoses as to where student(s) are currently understanding, choosing between multiple Interventions to implement, and Evaluating these decisions and thence continuing this cycle of DIE.

13. Students who share a classroom experience can have very different learning opportunities and experiences.

14. The importance of peer effects on students learning (Hattie, 2007).

15. Impact on teaching or learning. Perhaps the greatest issue is the lack of a corpus of knowledge about how observation changes either teacher practices or student learning. There is an emerging literature showing how multiple methods of teachers relate to each other. There is moderate correlations between observation, value-added, methods (Bell et al., 2012; Hill, Kapitula, & Umland, 2011; Kane & Staiger, 2012; Schacter & Thum, 2004) – but surely the point is the differential information these multiple methods bring to the overall question of effectiveness.

The conclusion seems simple: It is a sin to observe the teacher, but instead we need to observe the impact on the teaching on the students learning.

The question should be how to develop observations of teacher impact on students. For example, we (Popham and Hattie) are trying to create a brief schedule to use for this purpose (see below). Millar (2006) is using
spy-glasses (with video and sound) on students so that teachers can then see and listen what the students are looking at, and doing. Nuthal and Alton-Lee (2006) spent many years putting microphones on students and then testing various theories about how they engaged in learning, and the effect of teachers on this learning. There are also many video studies (perhaps the most famous are Anderson, 2010; Stigler, 2002) but more for describing classes than for evaluating teachers.

The Visible Classroom

Over the past 3 years, Clinton et al. (2014) in conjunction with Ai Media, and with implementation partner NESTA (and funded by Education Endowment Foundation), have developed a system utilising real-time speech-to-text, automated teaching analytics, and manual in-depth coding to support teacher reflection and practice development, and student engagement in the learning process.

At the heart of the project is the production of verbatim captioning in real-time, as well as transcriptions and analyses of full lessons, providing teachers and students with extensive feedback through a dashboard that illustrates various features of the lesson. This project grew out of a study which trialled the applications of the technology in providing greater access to classroom instruction and discussion for deaf and hard of hearing students in secondary settings (Clinton et al., 2013).

Providing useful feedback for teachers in real time, based on sound educational pedagogy, fits with models of best teaching practice like Visible Learning (Hattie, 2008). In Visible Learning it was noted that teaching and learning is too often hidden, with teacher talk often dominated by war stories rather than an evidence-informed view of learning. The impact of this technology is in making the content of the lesson really explicit and encouraging teachers to critically assess what they have
done and what their students have learned as a consequence. There is also a focus on development of collaborative teams such that teachers use the transcripts as a basis to initiate discussions with their colleagues around teaching methods and impact, and to facilitate them to work with colleagues to identify strategies for improvement.

At the end of each lesson, students in groups of 3-4 together rate the lesson on 7 dimensions. These are fed back (anonymously) to teachers and related to various aspects of their teaching over time.

As well, teachers receive analytics of their lessons.

This method has many advantages:

a. Feedback is immediate (within 7 seconds) and can be used by teachers and students
b. The analytics are immediate (1 hour) and are aimed at informing teachers as to how students “saw” the lesson, and the attributes of their lessons that relate to these outcomes
c. We are incorporating artefacts of students work to code for achievement (including progress, surface and deep).
d. We have developed “simpler” teacher talk (trialling with autistic students).
e. We are developing norms, aiming for 10,000 transcripts to be part of a researchable data base
f. We are aiming to “computerise” or at least use reasonable costing to code the lessons in an immediate sense (e.g., via content domains, age, etc.).
g. Unlike video, the transcripts can be quickly scrutinised, cut and paste (e.g., for use with para-professionals to revisit lessons with specific students), and cheaply stored.
h. Can provide examples of excellence of teaching, student voice, and artefacts
i. Can provide crowd sourcing to add to the coding, analytics, and understanding of lessons

Most important, our early work shows that teachers actually change their classroom behaviours to be more in line with student’s beliefs about enhancing their learning.

The early evidence shows that, even after 6 weeks, it is possible to see teachers changing their classroom practices to become more aligned with student experiences of learning.
Conclusions

Classroom observation of teachers’ impact on their students, augmented with teachers claims about how students are learning in light of their practices and decisions can become a worthwhile part of enhancing the quality of teaching – and this information can be used, in part, to understand the impact of teachers. But the current methods of observation schedules are too molecular, too unreliable, too costly, and not demonstrated to be formative for teachers or sufficiently related to students learning. Videos of the impact of teachers (by focusing on the students and their reactions and work rather than on teachers delivering and commenting) can be more powerful but are expensive, cumbersome, and often have major delays in reporting information that has been powerful for research on classrooms but less so for formatively enhancing teaching.

A more powerful method may involve captioning the teaching, involving the students in their evaluation of their learning, may provide a more exciting direction for “observing teaching”. It also fits with the premise of Visible Learning:
Evaluating and supporting educators

It is difficult these days to find a significant teacher-evaluation system in which a prominent kind of evaluative evidence is not classroom-observation data. Such evidence is typically collected by trained evaluators who observe classroom events to see if teachers (or students) engage in positive, research-supported actions while refraining from negative, research-admonished actions. In an effort to comply with federal preferences for “multiple-measures” evaluations, state officials are currently wrestling with the thorny problem of how to collect and use evidence of student growth. Even so, most states’ new or revised teacher-evaluation systems continue to employ classroom-observation evidence in their evaluation systems. In this proposed, designedly interactive symposium (featuring both small-group and total-group deliberations), three presenters, all of whom have given substantial attention to the role of classroom observations in the evaluation of teachers, will supply their views of the legitimacy and collection of classroom-observation evidence in the evaluation of teachers.

ABSTRACT:
It is impossible these days to find a significant teacher-evaluation system in which a prominent kind of evaluative evidence is not classroom-observation data. Such evidence is typically collected by trained evaluators who observe classroom events to see if teachers (or students) engage in positive, research-supported actions while refraining from negative, research-admonished actions.

The nature of U.S. teacher evaluation, of course, is fundamentally changing as a consequence of two influential federal initiatives, the 2009 Race to the Top Program and the 2011 ESEA Flexibility Program. Both of those programs call for states to establish more rigorous teacher evaluation programs in which multiple sources of evaluative evidence are to be employed and, among such evidence, student growth (typically students’ test performances) is to be a “significant factor.” In an effort to comply with these federal multiple-measures preferences, state officials are currently wrestling with the thorny problem of how to collect and use evidence of student growth. Even so, most states’ new or revised teacher-evaluation systems continue to employ classroom-observation evidence in their evaluation systems.

In this proposed, designedly interactive symposium (featuring small-group and total-group deliberations), three presenters who have given substantial attention to the proper role of classroom observations in the evaluation of teachers, will supply their views of the legitimacy of classroom-observation evidence in the evaluation of teachers. If classroom observations are to be employed in a teacher-evaluation program, each presenter will suggest how to maximize the quality of such evidence.

Presenters:

The initial presentation will be made by W. James Popham of UCLA. In a presentation entitled, Classroom-Observation Evidence in Teacher Evaluation: Playing the Ponies, he will argue that because evaluative judgments based on classroom observation data are, at best, probabilistic estimates of a teacher’s quality, the evaluative significance of classroom observation data should be dwarfed by evidence of student growth.

Scott Marion of the Center for Assessment will make the second presentation on the topic, Thinking Like an Evaluator: Improving the Validity and Reliability of Measures of Professional Practice. He will point out that recent studies, such as the Gates-funded Measures of Effective Teaching (MET, 2013), have raised concerns about the capacity of observers to accurately judge differences in instructional quality. He will also note that He will discuss concerns about the “thin sampling” issue associated with classroom observations. Marion will argue, therefore, that because classroom observations are not sufficient for characterizing professional practice, inferences about teacher practices must be supported by other means such as professional portfolios and a systematic collection and analysis of relevant artifacts.

John Hattie of the University of Melbourne will provide the symposium’s final presentation. Drawing on insights gained from his comprehensive review of more than 900 meta-analyses dealing with students’ learning, Hattie will raise a pivotal question regarding the evaluative use of classroom-observation evidence in his presentation, namely, Classroom-Observation Evidence: Looking at a Teacher or at a Teacher’s Impact?

Lou Fabrizio of the North Carolina Department of Public Instruction will moderate the small-group and total-group discussions. Substantial interaction among attendees is foreseen.