Symposium: Evaluating Teaching Quality

July 2019

Colin Penfold, Education Development Trust
Ann Childs, University of Oxford

Acknowledgements

The authors would like to extend their particular thanks to:

Emma Gibbs (Education Development Trust) for managing and reporting on the symposium, and also Astrid Fieldsend for transcribing the recordings of the discussions.

Diane Mayer, University of Oxford, for chairing the symposium.

The participants who presented the challenges, namely:

- Daniel Muijs, Ofsted
- Katharine Burn, University of Oxford
- Colin Penfold, Education Development Trust
- Ian Thompson, University of Oxford
- Trevor Mutton, University of Oxford
- Teresa Tatto, Arizona State University

All the participants who took part in the symposium in person or online, for sharing their knowledge and expertise, whose valuable and challenging contributions to the discussions are the basis of this proceedings paper.

Participant list

Charlotte Bergin	Inter-Agency Network for Education in Emergencies
Barbara Bruns	Centre for Global Development
Katharine Burn	University of Oxford
Ann Childs	University of Oxford
Maria Cunningham	Teacher Development Trust
Emma Gibbs	Education Development Trust
Jenny Gore	University of Newcastle (New South Wales, Australia)
Richard Graham	Ark (Education Partnership Group)
Jenni Ingram	University of Oxford
Sharon Kim	New York University (TIPPS)
Richard King	Education Development Trust
Ariel Lindorff	University of Oxford
Diane Mayer	University of Oxford
Tony McAleavy	Education Development Trust
Cheryl McGechie	Education Development Trust
Daniel Muijs	Ofsted
Trevor Mutton	University of Oxford
Colin Penfold	Education Development Trust
Mahjabeen Raza	New York University (TIPPS)
Anna Riggall	Education Development Trust
Pam Sammons	University of Oxford
Lorena Sernett	Teachstone
Iram Siraj	University of Oxford
Teresa Tatto	Arizona State University
Ian Thompson	University of Oxford
Sam Twiselton	Sheffield Hallam University
Emily Woolf	Department for International Development

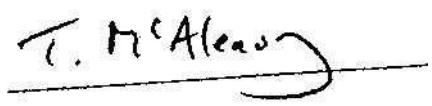
Foreword

I am delighted to endorse this report, which sets out the proceedings from a symposium on the evaluation of teaching quality. The symposium was organised in July 2019 by Education Development Trust, in partnership with Oxford University's Department of Education. I would like to thank my colleagues and the staff from Oxford for planning and facilitating such a successful event. We have a long tradition of partnership with the University of Oxford and this symposium further built upon this successful relationship.

Teaching quality is immensely important to education reform around the world. We are witnessing a worldwide learning crisis and effective responses will depend upon our capacity to improve teaching on a global scale. Any attempts to transform teaching quality depend, as an essential precondition, on the accurate measurement of that quality. Therefore, as an education research and consultancy organisation, Education Development Trust's interest in teaching quality – and how to measure it – is a central concern and key priority.

We have the privilege of operating both in the Global North (particularly the UK) and the Global South (particularly Africa). We are keen to promote an inclusive global dialogue about teaching quality, while recognising huge contextual differences from place to place. It was therefore particularly pleasing to have a wide range of professional and geographic perspectives as part of the symposium, identifying both common issues and specific challenges in the evaluation of teaching quality worldwide.

The symposium was structured around the discussion of several issues associated with evaluating teaching quality, with each issue addressed in the form of a brief initial challenge from one of the participants, followed by debate and discussion. This format worked brilliantly. The following pages detail some of the key themes, opportunities and challenges that were highlighted during the event. For me, this represents a significant contribution to the discourse on the evaluation of teaching quality, and I hope that the rich dialogue generated at this event will continue in the coming years.



Tony McAleavy
Research Director
Education Development Trust

Introduction

Diane Mayer, Professor of Education (Teacher Education), University of Oxford

Issues of teacher quality are being examined around the globe, often prompted by international assessments and related country comparisons, as well as the complex demands of teaching an increasingly diverse student cohort within increasingly diverse communities. As a result, policies are being developed and reform agendas are being implemented with a view to improving teaching quality and learning outcomes for all students. Such agendas often involve increasingly complex systems of regulation and accountability, with statements of professional standards and ways of making judgements against these standards. At the point of initial teacher training, this includes capstone performance assessments designed to demonstrate that graduates have the necessary capabilities for beginning teaching. Judgements are also made after a period of induction, usually to determine full credentialing or registration. Ongoing assessment and demonstration of sustained and even increasing effectiveness at subsequent stages of one's teaching career are the subject of ongoing and sometimes heated industrial and professional debates.

While thinking about the teaching career trajectory in these ways focusses attention on evaluating teaching for accountability-related purposes, we might also consider the impact these measures have on teachers' professional knowledge, practices and learning, beyond simply rating a teacher against predetermined criteria. Even if we only frame the purposes of evaluating teaching in these ways, research shows that being involved in the evaluation of their own teaching – especially if they have some input into what is being judged, how, and when – positions teachers well to reflect on and inquire into their own problems of practice, thus supporting their professional growth. In addition, of course, there are opportunities for peer and within-school evaluations designed to facilitate professional learning, often in collaborative ways.

Irrespective of purpose – but certainly informed by purpose – there are decisions to be made about the methods used for evaluating teaching, the criteria by which teaching will be evaluated, and who will conduct these evaluations. Moreover, issues associated with varied teaching and learning contexts, as well as scalability of approach, need to be considered. Therefore, this symposium was very timely, providing an opportunity to have a series of discussions around these important issues. By examining them from different points of view and highlighting points of agreement and disagreement, gaps in the current knowledge base can be identified to inform future research.

Thank you to Colin Penfold from Education Development Trust for instigating this symposium, and to Ann Childs and her colleagues from the University of Oxford for supporting Colin and ensuring the symposium was such a success.

Overview of the symposium

The symposium was a joint endeavour by Education Development Trust and the Department of Education at the University of Oxford.

It took place on 23 July 2019 at the Department of Education, University of Oxford. Most participants attended in person, but some attended online, as they were in different parts of the world.

The symposium was structured around six challenges associated with evaluating teaching quality. Each challenge was presented in the form of a brief (approximately five minute) 'provocative' presentation from one of the participants, followed by around 30 minutes of debate and discussion amongst the group. The aim was not to necessarily come to agreement, but to discuss the full range of issues, raising different sides of the argument. The symposium was chaired by Diane Mayer, Professor of Education (Teacher Education) at the University of Oxford.

The six challenges were:

Challenge 1:	Is pedagogy the right thing to look at if we want to understand the quality of education pupils receive?
Challenge 2:	How effectively can we evaluate teaching quality if we only focus on generic pedagogy and do not consider the subject being taught?
Challenge 3:	Can observations of teachers by non-experts be as valid and as useful as those carried out by experts?
Challenge 4:	Are snapshot lesson evaluations as valid and useful as observations of full lessons?
Challenge 5:	Is it realistic for evaluation of teaching to be a teacher developmental process?
Challenge 6:	Can we evaluate teaching quality fairly without taking into account the context and previous experiences that teachers have had, including their knowledge and beliefs?

During a final session at the end of the symposium, participants were invited to raise any challenges or issues that had not been discussed, and also to identify possible actions moving forward.

This paper was compiled from notes taken during the meeting and a recording of the discussions. As a result, the report has, at times, a more informal tone which captures the richness of the original challenges and the wide-ranging and diverse discussions which followed.

Challenge 1: Is pedagogy the right thing to look at if we want to understand the quality of education students receive?

Initial provocation: Daniel Muijs, Head of Research, Ofsted

There is a long tradition of research on pedagogy and teacher effectiveness which has contributed to what we now know about teaching. This has in turn contributed to teacher education and professional development. However, the focus of this first provocation is: are we focussing on pedagogy too much? And have we reached the end of the line of what we can look at in terms of pedagogy?

There are four major reasons for which we need to ask these questions.

Firstly, focussing on pedagogy, particularly when we look at the way teachers teach, is quite limiting in terms of what we can find out about how well teachers actually know their subjects. For example, there are many instances of teachers being observed where they have very good pedagogical skills and a nice repertoire but still lack subject knowledge. In one particular case, Ofsted carried out research in a primary school, where the researchers saw a very engaged group of pupils doing an activity where they were making Viking ships. While the children were all obviously enjoying the task, when the researchers asked them, 'Why did the Vikings need ships?', the children were unable to answer the question, as this has not been taught by the teacher. It is possible to look at a lesson but still miss lots of relevant information, and it is possible to be able to teach well, but still teach content poorly.

Secondly, related to this first point, it is conceivable that – in addition to the possibility of teachers omitting content – they can also disseminate harmful content. For example, a teacher may be highly effective at teaching but may actually teaching intolerance or prejudice. This creates a risk of children exiting the school system with those beliefs. They may have learned the concepts well, but there is a negative societal impact. Again, it is not just about **how** teachers teach, but **what** they teach.

The third point is that, by looking at individual lessons, we do not gain an understanding of the sequence of teaching: a lesson in isolation does not necessarily mean very much without an understanding of where it sits in a sequence of lessons on a particular topic. That can lead to misconceptions and mistaken conclusions. For example, an observer can go into a lesson and see that pupils are spending some time working individually on a set of worksheets. Depending on the framework being used, the observer may say that there is not very good teaching going on – there is no interaction with pupils, etc. However, this does not consider context. It may be that, within the overall sequence of lessons, this is exactly the right activity for the pupils to be doing at that particular moment. Therefore, it is difficult to understand pedagogy without looking at it across the sequence.

Finally, pedagogy is only a small part of what happens in a school and students' wider learning. There is a whole other set of factors as well, including what happens in the wider school environment, and what happens outside of school. For example, many children get significant input into their learning outside of school (from private tutoring etc.). Therefore, we need to look at quality of education as something that is much broader than what happens in the classroom.

To summarise, pedagogy is important and all the research on pedagogy is very valuable. However, it leaves us significantly short of understanding learning, development and quality of education, if we don't contemplate **first and foremost** the **content** that is actually being taught.

Discussion

A participant asked Daniel to elaborate on his point relating to teaching 'harmful content', and to clarify whether this is done on purpose or by accident. Daniel replied that it can be either: in some inspections, there have been deliberate attempts to teach students harmful content, for example, extremist views. In some cases, however, it is absolutely not deliberate and stems from limitations in a teacher's own understanding.

Another participant explained that what she took from this challenge was that it is important to look at a variety of things when trying to understand the quality of education and that one needs to place the snapshot of evidence into a broader understanding of the context. Daniel confirmed this was the case.

A further participant stated that previous research (particularly within teacher and school effectiveness traditions) had found that good content knowledge is a necessary condition – but not a sufficient condition – for effective teaching to be considered high quality. There is some research which suggests that at least three to five observations are needed to get a good understanding of a teacher's practice, i.e. a longer-term sequential approach is preferable. Furthermore, it is important to be cognisant of the fact students have multiple teachers, including out-of-school teachers (private tutoring). It is known that private tutoring can have a measurable effect on student outcomes. Therefore, it is necessary to understand the wider context, as well as the culture of the school. If students have multiple teachers, it stands to reason that consistency in teaching approach, alongside teachers collaborating and working on professional development, leads to a positive set of experiences for students (across different teachers, across time, and across schools). Daniel asked which is the most important, based on this thinking: teachers being consistent, or the actual pedagogy used? The participant felt that the two were not mutually exclusive and that both factors could be equally important.

A participant asked whether testing was better than observation as a way of assessing the quality of learning, since testing has the capacity to manage some of the problems identified. Daniel stated that testing itself has major limitations, specifically, that it can only ever provide a snapshot of what has been learned, and that it is only ever possible to test a small portion of a topic. Therefore, testing does not have a simple relationship to quality of education. Test preparation methods can skew results and results are strongly influenced by student background. It is therefore important to be cautious, but a combination of test results and observations could be an appropriate means of assessment.

Another participant raised the point that content knowledge used to be considered very important, and this led to an overemphasis on it, particularly at secondary level. It is therefore crucial to keep the emphasis on **how** content is delivered, not just what the content is. She suggested that pedagogy was one side of a coin, with the other side being curriculum, and that the two sides were held together by assessment. Therefore, observations which do not take into consideration teachers' planning, teachers' intentions, or what the teachers have done before are clearly failing in their evaluation of the teaching. She argued that snapshot observations might be of different lengths of time, depending on the context.

Daniel's response was that pedagogy is **how** you teach something, the curriculum is **what** is taught, and assessment is a way to measure the impact. It is not possible for these to be independent of each other as they have a close relationship. However, Daniel disagreed that they are two sides of the same coin. For example, the Core Knowledge curriculum developed by E. D. Hirsch demands that it be delivered via direct instruction. However, an evaluation carried

out by Johns Hopkins University[1] found that many schools delivered the Core Knowledge curriculum using people-centred pedagogical approaches. This shows that, even when there is supposed to be a very strong relationship between pedagogy and curriculum, in practice, this is not always the case.

Another participant stated that in her work she had found that pedagogy is too frequently understood to be the 'how'. The view of pedagogy she had found most helpful was David Lusted's work from 1986[2], in which he writes that pedagogy as a concept draws attention to the process through which knowledge is produced. He suggests that pedagogy addresses the 'how' questions involved not only in the transmission or reproduction of knowledge, but also in its production. Indeed, it enables us to question the validity of separating these activities so easily, by questioning the conditions under which and the means through which we come to know. How one teaches becomes inseparable from what has been taught and, crucially, how one learns. The participant agreed with Daniel's perspective that pedagogy and curriculum are not the same thing and that notion of inseparability is critical in how pedagogy is understood. The participant described the work she has been doing for a couple of decades that led to the 'Quality Teaching Model'[3] (QTM), which is a lens through which to evaluate pedagogy. It has three dimensions: intellectual quality, quality learning environment and significance. It deals with how knowledge, students and issues of equity are treated. These questions about how pedagogy is conceptualised are really important. It is important to consider context when thinking about the quality of pedagogy or teaching. Having a framework that is adaptable to different contexts – but also starts to codify what good teaching looks like – is really important.

Another participant summarised the discussion so far into two key questions:

1. Is pedagogy *enough* to understand quality of education?
2. What do we mean by pedagogy?

The participant went on to suggest that traditional definitions of pedagogy have very much tended to focus on 'doing', but he argued that pedagogy is also about what teachers know and what teachers believe.

Another participant raised a point related to a large-scale piece of research (across 17 countries) which encountered great difficulty in getting agreement on what pedagogy means. Consequently, pedagogy was very difficult to measure. The participant stressed the need for countries to establish common definitions in order to develop meaningful measures. In the research project described, Shulman's typology[4] helped with the development of a common definition.

A participant then raised the point that there will be different pedagogical strategies for different groups of students and at different stages. For example, disadvantaged students may require more scaffolding. What is perhaps less context-dependent is subject knowledge: there should not be lower expectations of subject knowledge for disadvantaged groups or different parts of the world. The best teachers are needed in the most disadvantaged contexts. Being an excellent teacher is also about other learning outcomes beyond exam performance: motivation and engagement, student self-efficacy and student independence.

---

[1] Stringfield, S., Datnow, A., Borman, G., and Rachuba, L. (2000). National evaluation of Core Knowledge Sequence implementation: Final report. Baltimore, MD: Johns Hopkins University, Center for Research on the Education of Students Placed At Risk

[2] Lusted, D. (1986). Why Pedagogy? Screen, 27(5), 2-14

[3] https://theelements.schools.nsw.gov.au/introduction-to-the-elements/policy-reforms-and-focus-areas/quality-teaching-framework.html

[4] Shulman, L. S. (1987). Knowledge and teaching: foundations of the new reform, Harvard Educational Review, 57(1), 1-22

Another participant wanted to build on the point about scaffolding, stating that especially in low-income settings, contextual factors around teacher motivation and accountability become critical. For example, Zambian community schools, largely staffed by non-trained community volunteers, outperformed state schools where teachers had received trained.[5] The theory behind this was that teachers coming from within the community had more intrinsic motivation, and that they felt more accountable to the community. This led to lower levels of staff absenteeism and greater efforts by the teachers to do their best for the students. This counted for more than the three years other teachers had spent at teacher training college.

One participant talked about observation and how one's background and beliefs affect what one sees as an observer. They argued that it is therefore more reliable to observe lessons in collaboration with others. It was also noted that students and teachers change their behaviours when they are observed. Therefore, what is being seen is not typical of the education that is delivered or received.

Another participant talked about how different approaches to evaluating pedagogy are more or less relevant at different times. She spoke about her experience working with the Ministry of Education in Nigeria, which introduced a test for teachers' content knowledge. As time passed, the ministry began to use other measures to tell them about teacher quality. There is an argument for not necessarily using all possible methods for evaluating pedagogy at the same time, but rather thinking about how these evaluations could be sequenced in relation to the development journey of the education system. Another participant agreed, suggesting that it is often not practical to simultaneously adopt different approaches to evaluating pedagogy to build the 'perfect picture' of teaching quality. Indeed, there is no one correct methodology, and there must be an acceptance that a variety of considerations need to be made in order to arrive at a valid judgement.

Another participant felt it was important to note that we 'look' at teacher quality for different purposes. If the purpose is to develop practice and provide formative feedback, then what we understand by pedagogy becomes very important – as do other factors, such as teacher beliefs and teacher efficacy. The participant suggested that Gage's views on the science and art of teaching[6] were relevant. The science is the theoretical understanding of pedagogy and the art is the linking with the context and with reality.

In closing, Daniel acknowledged the usefulness of the discussion. He felt, however, that the discussion had moved away from the original challenge. He stated it was interesting to hear the difficulties of defining pedagogy across different contexts. A general weakness in research is the lack of clarity we have around concepts and this leads to the danger of 'conceptual creep'.

[5] https://www.epdc.org/sites/default/files/documents/Zambia%20Community%20Schools.pdf

[6] Gage, N. L. (1978). The scientific basis of the art of teaching. New York, New York: Teachers' College Press

Challenge 2: How effectively can we evaluate teaching quality if we only focus on generic pedagogy and don't consider the subject being taught?

Initial provocation: Katharine Burn, Associate Professor of Education, University of Oxford

My starting point in addressing this question is undoubtedly conditioned by the fact that one of my roles is that of history curriculum tutor within the University of Oxford's Internship Scheme: a secondary programme characterised not only by its long-standing and close partnership with local schools, but also by the strength of its commitment to subject-specific professional preparation. It is the curriculum programme, jointly planned by school-based subject mentors and university-based curriculum tutors, that lies at the heart of the interns' experience and on which most attention is focused. Although only one of the eight Teachers' Standards (DFE, 2011)[7] is explicitly concerned with subject – and even that is expressed in terms of a narrow emphasis on subject and curricular knowledge (as opposed to subject-specific pedagogy) – it is the judgement of the subject mentor and curriculum tutor that essentially determines the award of Qualified Teacher Status.

Yet, even in a course such as this, structured around the subject, with half of the new mentor induction programme and all subsequent mentor meetings held in subject groups, it is repeatedly necessary to reiterate the importance of the subject dimension because of the way in which some mentors (or other subject teachers with whom the interns are working) tend to neglect it as they observe and review lessons with the interns. Indeed, a year ago, the decision was taken to modify the regular lesson-observation pro-forma to ensure that in all cases at least some of the observer's attention was focused on issues specific to subject teaching and learning. The fact that the team (collectively) judged this to be necessary suggests that not all school-based mentors share the strong convictions of their university partners that effective evaluation and appropriate advice about further development require explicit attention to the subject being taught.

It would appear that Ofsted, the body responsible for inspection of the quality of teaching in state-maintained schools in England, may be similarly unconvinced about the value – or at least the necessity – of a subject-specific focus. The seminar recently convened by Ofsted that reviewed six different models of lesson observation to gain an 'international perspective' that would inform future development of their own inspection framework[8] included only one subject-specific model: Mathematical Quality of Instruction (MQI)[9].The other models[10] all adopted a generic approach.

[7] DFE (2011). Teachers' Standards Guidance for school leaders, school staff and governing bodies. Department for Education. Available online at: https://www.gov.uk/government/publications/teachers-standards

[8] Ofsted (2018). Six models of lesson observation: an international perspective. Report 180022. Available online at: http://www.sici-inspectorates.eu/getattachment/7e62a765-f133-48c8-bb16-e7f76b2829f3

[9] Charalambos, C.Y. and Litke, E. (2018). Studying instructional quality by using a content-specific lens: the case of the Mathematical Quality of Instruction framework. ZDM Mathematics Education 50(3), 445-460

[10] The other models considered were: the Classroom Assessment Scoring System (CLASS); the Framework for Teaching (FfT); the International Comparative Analysis of Learning and Teaching (ICALT); the International System for Teacher Observation and Feedback (ISTOF); and the Generic Dimensions of Teacher Quality model.

In reflecting on the similarities and differences between the models, Ofsted (2018) observed that:

> 'The structure of the six models were typically informed by the research literature on the quality of teaching. This explains some of the overlap noted between the models, particularly around their focus on classroom management, instruction, student behaviour and attitudes.' (Ofsted, 2018, p.6)

While the first and third of these elements (classroom management and student behaviour and attitudes) could plausibly be evaluated without reference to the subject dimension, it is difficult to conceive of a focus on instruction that did not, at least implicitly, require consideration of those dimensions of practice that depend on the knowledge base that Shulman (1986)[11] originally identified as 'pedagogical content knowledge'. If we examine the five 'measurable domains' within the MQI model specifically focused on subject – in this case, mathematics – they reflect an intention to capture the nature and quality of the mathematical content that is available to students, as expressed through teacher-student interactions, teacher-content interactions and student-content interactions. These domains are:

- Common core-aligned student practices (captures the ways in which students engage with mathematical content)
- Working with students and mathematics (identifies whether teachers can hear and understand what students are saying mathematically and respond appropriately)
- Richness of mathematics (measures the attention to the meaning of mathematical facts and the procedures and engagement with mathematical practices and language)
- Errors and imprecision (identifies mathematical errors and distortion of content by the teacher)
- Classroom work is connected to mathematics (captures whether classroom work has a mathematical point or whether instructional time is spent on activities that do not develop mathematical ideas).

(Ofsted 2018, p.18)

Consideration of each of these domains readily reveals how much detail about the nature of the three different kinds of interaction is likely to be lost without an explicit focus on the subject. Similar reflection on the level of observer knowledge required to judge the quality of teachers' practice and students' learning in relation to the different interactions immediately highlights the challenges associated with undertaking such evaluation. Those challenges become self-evident once we consider the initial training and ongoing monitoring required for observers to use the MQI instrument. This involves observation (usually video-based) of each seven-and-a-half-minute lesson segment by two observers, both with strong mathematical knowledge, trained and certified online and supervised weekly. The costs and practical difficulties in sourcing potential observers with the necessary knowledge base, even before providing the necessary training and supervision, clearly place the use of subject-specific tools such as MQI beyond the realms of what is feasible in most contexts.

In exploring the scope for some kind of 'middle way' that recognises the importance of the subject domain while acknowledging the principles that are fundamental to the quality of teaching in all domains, my attention has been drawn recently to two models being used as professional development tools, rather than as observation instruments: the collection of 'high leverage

[11] Shulman, L. (1986). Those who understand: knowledge growth in teaching. Educational Researcher, 15(2), 4-14

practices' (outlined, for example on the 'TeachingWorks' website, under the auspices of the University of Michigan[12]) and the 'Quality Teaching Model' developed and refined in New South Wales (Australia) that is intended to be used by teachers as a reflective tool.[13]

While the main impetus behind the first of these – the attempt to identify a set of core practices that had the 'greatest impact on student learning' – was the hope that such a list could serve as a curricular framework for the professional preparation of beginning teachers, my own encounter with 'core practices' was in the context of a professional development institute for experienced teachers, offered by the Stanford Centre for Excellence in Teaching. What was perhaps most striking about the programme and the way in which participants engaged with it was the fact that while teachers from three very different national contexts (Australia, America and Hong Kong) could work very productively together in discussion and experimentation or role-play with the practices, that collaboration took place within subject groups. In deciding that all the detailed developmental work should take place within subject domains, the course leaders were, of course, reflecting the way in which the 'high leverage' or 'core practices' research and development traditions, as co-ordinated at Ball and Forzani (2011)[14] at the University of Michigan, or by Pam Grossman (2018)[15] at the University of Pennsylvania, have respected the nature and demands of particular subjects. Additionally, they reflected the way in which pedagogical practices within each domain have to be not simply applied to relevant content knowledge but richly informed by relevant pedagogical content knowledge. The latter does not simply involve an understanding of how the content can be broken down and made accessible to diverse learners; it also encompasses an appreciation of how most students think about, make sense of and use the subject.

In contrast, one of the main messages that I took from engaging in a condensed version of the training provided for teachers leading professional development using Quality Teaching Rounds, was the insistence of its principal architect, Jenny Gore, that the model worked most effectively with teachers of different subjects working together (ideally in groups of four). Its approach – using a carefully structured collaborative, and essentially democratic, cycle of lesson observation and discussion – was, she suggested, capable of bridging subject domains. In response to questions about how generally applicable it might be, Gore also claimed that the way in which identification of the 18 dimensions of the Quality Teaching Model had drawn so widely on different research traditions, also meant that it could be used effectively to drive improvement, regardless of the particular philosophy of teaching (adherence to 'traditional' or 'progressive' pedagogies, for example) embraced by the teachers or the systems within which they worked. This commitment to cross-curricular working did not, however, amount to a claim that subject does not matter. The first dimension of the Quality Teaching Model, 'Intellectual Quality', includes three elements that focus attention very directly on subject. The distinction between 'Deep Knowledge' (what is brought into the lesson) and 'Deep Understanding' (what students take away from the lesson) seems to be sufficient to focus the attention of both the teacher and the observer on the processes and achievement of subject learning. Meanwhile, the dimension of 'Problematic Knowledge' helps to raise important questions about the nature and status of knowledge within the domain and whose knowledge counts. While the promise of this model, at a relatively low price (dependent on two days' training for participants from each school, and then measured in terms of the time needed for shared observation and discussion), is only now being

[12] http://www.teachingworks.org/work-of-teaching/high-leverage-practices

[13] Bowe, J. and Gore, J. (2017). Reassembling teacher professional development: the case for Quality Teaching Rounds, Teachers and Teaching, 23(3), 352-366

[14] Ball, D. L. and Forzani, F. M. (2011). Building a common core for learning to teach, and connecting professional learning to practice. American Educator, 35(2), 17–21, 38–39

[15] Grossman, P. (2018). Teaching Core Practices in Teacher Education. Harvard, MA: Harvard Education Press

tested beyond New South Wales, in a large-scale randomised control trial[16], its potential certainly seems worth exploring.

It is important to note, however, that the architect of Quality Teaching Rounds has also categorically insisted that the Quality Teaching Model and scoring rubric should be used for developmental rather than assessment purposes, arguing that its power lies in the questions that the tool prompts teachers and their peers to ask about their own practice, not in the judgments that it allows others to make about that practice. Given all that is known about effective formative assessment depending on learners' own grasp of the assessment criteria, should we perhaps take heed of Gore's message and recognise that the best way of trying to tackle the subject dimension might be to focus our limited resources on ways of making the essential criteria explicit to teachers and training them in their use, rather than training the observers?

Discussion

A participant explained that her organisation is frequently involved in supporting groups of teachers to improve, and helping governments to support large groups of schools and/or teachers to improve. She asked the question, which should come first: the theory about what teachers should be doing better or the theory around and the tools that allow us to observe this? She continued by agreeing with the idea that subject expertise is important. However, from her experience running the Organisation for Economic Co-operation and Development (OECD) Teaching and Learning International Survey (TALIS) video study, which focused on a very specific area of mathematics, she explained that observation was both exciting and complex. It would not be possible to do what is being done in TALIS in the field. She asked about what tools would be needed on the research side to make this possible. Katharine responded by saying that the subject itself is part of the answer: what does learning, understanding, and producing knowledge in this subject look like?

Another participant suggested that looking at this issue from a systems perspective (e.g. standards and data) might produce a different answer. Another participant agreed that different approaches are needed depending on the purpose of the observation. Conflating approaches used for performance management with professional development can drive behaviours that are unhelpful. He agreed with the point made during the presentation that the model described should be used for teachers' own development and, if for system-wide interventions, alternative tools should be considered. He argued that looking from the systems perspective, there are basics that need to be in place before focusing on subject. For example, it is necessary to establish the starting point of the system that is being investigated. If it is a system where teachers fail to turn up to school, then this is the starting point. This was supported by another participant who talked about a study which tested the mathematical competence of primary school teachers in southern African countries. It found that the vast majority of teachers had not mastered the fourth-grade mathematics curriculum. Although it would be possible to train these teachers in generic pedagogical approaches, the impact would be limited due to their lack of competence in the subject.

Another participant agreed that the subject was very important. However, she pointed out that in Initial Teacher Training (ITT), research has shown that teachers need different things at different times, especially when recognising that teachers will be in schools in different contexts. If, for example, a school has no behaviour system, it is may be difficult for a teacher to move beyond dealing with managing behaviour to focus on the subject. Therefore,

[16] https://www.newcastle.edu.au/research-and-innovation/centre/teachers-and-teaching/quality-teaching-rounds/building-capacity-for-quality-teaching-in-australian-schools

the tools and instruments used to evaluate teaching and support teachers to improve should reflect their different needs.

A participant stated that no one measure or one source of evidence is enough to make high-stakes judgements about teachers and that this should be accepted as the starting premise. When using teacher value-added data, the numbers are too small for sufficient confidence intervals. Observational data alone is weak because of the number of observations or sequences of observations required to build an accurate picture of teaching quality. Additionally, different observation frameworks might lead observers to different conclusions. It is therefore better to use this kind of data to focus on improvement and professional development. Other sources of information could support this, such as student perceptions of teaching, peer observations etc. This might need to be carried out in subject-specific teams, especially at secondary, but perhaps also at primary level. It should be acknowledged that as soon as classroom observation becomes high-stakes, teacher behaviour is altered.

Another participant built upon the previously raised ideas concerning teachers owning any evaluation process, rather than it being used as a high-stakes judgement of the teacher. She stated that if large-scale improvement is the goal, then working with and in support of teachers is critical, as opposed to setting up schemes that 'terrorise' them. The work done on Quality Teaching Rounds[17] in New South Wales, Australia, brings together teachers who work in different parts of the school. It focuses on generic teaching practice, although what is being taught and how it fits with the key concepts of the discipline are part of the conversation. The key finding of this methodology is that the process of bringing of different groups of people together to work in this way is a catalyst for driving improvement. A Randomised Control Trial (RCT) has shown a strong improvement in the quality of teaching and in teacher morale. There are three findings that have been drawn from this:

1. Bringing teachers together across diverse disciplinary and/or stage-/age-specific teaching backgrounds develops fresh insights about pedagogy and students;
2. Quality Teaching Round activity enhances collegiality across the school; and,
3. Levels of professional collaboration after participating in Quality Teaching Rounds increases.

The participant argued that when looking at system-wide change or attempting to bring about change at scale, generic frameworks have some advantages in terms of cost (developing detailed and specific frameworks is extremely costly, as well as complex). However, beyond this, teachers themselves are capable of generating change once they are introduced to the new concepts and the ways of working.

A participant reflected on earlier comments regarding the tools used to evaluate teaching quality, stating that several RCTs have been carried out which looked at Continuing Professional Development (CPD) rather than Initial Teacher Training (ITT). The significant finding from these trials was that a generic focus on teaching was particularly useful. However, it is important that the evaluation tools have a level of reliability. She also stated that such tools need to go beyond concurrent validity (comparing a new tool to an existing tool) to predictive validity (the degree to which the tool predicts future student outcomes). Testing observation tools is important if we are to have predictive validity: hundreds of

[17] https://education.nsw.gov.au/teaching-and-learning/professional-learning/scan/past-issues/vol-36,-2017/quality-teaching-in-our-schools

schools need to be looked at. A tool can be used to look at strengths and limitations of teaching practice and thus it can be used as a diagnostic tool. Where this has been done (evidence-based professional development), there have been observable improvements in child outcomes.

Another participant advocated helping teachers rather than 'terrorising them'. She spoke about a study she had conducted, in which the teacher educators of an initial teacher training course were asked to assess the trainee teachers in both content knowledge and pedagogical content knowledge. The teachers were not threatened at all: in fact, they were happy that people were working with them to 'help build the profession'. The participant felt that all the different approaches to evaluating teachers can work together. She agreed that teaching happens in the context of the subject matter, the context of the students being taught and the context of the teacher's capabilities. She proposed the idea of thinking carefully about the language around 'doing to' and 'doing with' teachers. She suggested that the reason teachers have been 'terrorised' by assessment is because they have been the subjects of assessment, rather than participants in assessment. She went on to state that it is often assumed that teachers already know what they need to know, but this is not necessarily the case, particularly in developing countries. Many teachers do not have the essential knowledge they need to be effective and, in many cases, they must rely on textbooks and learn along with their students.

Another participant agreed with the point about assumptions being made about teachers' knowledge base, especially in low- to middle-income countries. She gave an example of some work being done by the Inter-American Development Bank in Ecuador[18] using the CLASS[19] tool, which focused on generic teaching. There have been interesting results using the tool in early grades, which has found that good teaching crosses subjects: the study has not found that teachers are good at one subject and weak at another. But in the context of quality interactions, the study suggested that teachers were either good or they were not. This suggests that good teaching in the early grades is independent of subject matter.

Another participant agreed, noting that it is relatively easy to look at the difference between poor practice and average practice using generic instruments, and to pick up many of the big issues around factors such as climate and behaviour management. These kinds of generic instruments tend to be much more problematic when we want to go beyond that and look at the highest quality teaching. The participant argued that this is because in the highest quality teaching, subject knowledge and subject pedagogy becomes much more central.

Another participant argued that it is precisely because generic tools are more straightforward that we find them attractive. It takes considerable expertise to measure more complicated items. Furthermore, it is essential to move beyond thinking only about subject, but also about the nature of the subject: what it means to be a linguist or a historian. For example, modern foreign languages (MFL) has become a subject concerned with the controllable and measurable, rather than about communicative competence in the language.

In closing, Katharine suggested that the most prevalent idea in the discussion had centred around the notion that no one source is enough to determine teacher quality. She agreed with the ethical concern raised by some participants: if we are doing research, it should be

[18] For example: Caridad Araujo, M., Carneiro, P., Cruz-Aguayo, Y. and Schady, N. (2016). Teacher quality and learning outcomes in kindergarten. IZA Discussion Paper No. 9796. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2750279

[19] https://curry.virginia.edu/classroom-assessment-scoring-system

for the purpose of helping teachers. Evaluation tools should exist to help teachers rather than punish them. She expressed some concerns around the ideas of ‘stages’ of teachers’ learning (i.e. the idea that teachers can only think about behaviour when starting out). Even if they are not able to establish classroom climate immediately and this takes most of their focus, novice teachers will still be highly concerned with what is being taught. She reinforced one participant’s observation about the important relationship between pedagogical content knowledge and students’ understanding. She thought that the discussion on Quality Teaching Rounds, and how they have effective because they bring teachers together to learn about their students, was illuminating.

Challenge 3: Can observations of teachers by 'non-experts' be as valid as those carried out by 'experts'?

Initial provocation: Colin Penfold, Principal Consultant, Education Development Trust

In answering this question, there are two things that I want to address:

1. What I mean by 'expert' and 'expertise' when evaluating teaching; and,
2. Why I think that expertise matters.

The first thing to say, is that as noted by Evans and colleagues (2015)[20] there is little empirical research evidence on how effective non-specialists observe lessons compared to subject specialists. In their own small-scale study, they found that it was possible to distinguish between notes of observations made by specialist and non-specialist observers. This certainly relates to my own experiences when observing lessons with non-specialists. I suggest that there are important differences in what specialists and non-specialists observe.

So, what do I mean by 'expert'? I do not think that it is easy to define what an expert teacher is. Berliner[21] states that,

> '…expertise is specific to a domain and to particular contexts in domains and is developed over hundreds and thousands of hours.' (Berliner, 2004, p.201)

He goes on to say, that in relation to teaching,

> '…it is likely that almost every expert pedagogue has had extensive classroom experience.' (Berliner, 2004, p.201)

According to Berliner, there is a clear relationship between expertise and time (experience). However, experience by itself does not equal expertise. Expertise depends on the breadth and richness of experience. Furthermore, expertise also relates to domain-specific knowledge: with respect to teachers, this is the knowledge for teaching. Clearly, in relation to understanding what we mean by an 'expert teacher', we need to unpick what kind and length of experiences are necessary to become an expert, and also to identify the knowledge base for teaching that expert teachers not only possess but use.

What therefore do we mean by 'expert' observers of teaching? Clearly, this is about experience and about knowledge. I have not personally come across much that has been written about this, so the following attributes are my own initial thoughts and ideas. I suggest that expert observers of teaching:

1. Have the ability to use an observation framework or tool objectively and consistently. I have identified this first, not because I think that it is the most important attribute, but because it seems to be what many people focus on when they begin to think about identifying and training observers. Clearly, objectivity and consistency are critical traits of effective classroom observers. It appears to me that the current vogue, particularly in developing country contexts, is to take graduates, or whoever happens to be available, and then train and assess them in using an observation tool. The belief seems to be that this is not only necessary but also sufficient for

[20] Evans, S., Jones, I. and Dawson, C. (2015). Do subject specialists produce more useful feedback than non-specialists when observing mathematics lessons? Proceedings of the 38th Conference of the International Group for the Psychology of Mathematics Education, 3, 33-40

[21] Berliner, D. C. (2004). Describing the behavior and documenting the accomplishments of expert teachers. Bulletin of Science, Technology and Society, 24(3), 200-212

effective observation of teaching. The focus appears to be solely on rater reliability. I think that this argument is fatally flawed for several reasons. As Coe (2014)[22] reports, one of the findings of the Measures of Teacher Effectiveness (MET) project[23] was that observer training does not ensure quality observations. Although the ability to be objective and consistent is important and necessary, it is far from sufficient. I believe that observing teachers teach is a highly skilled activity. It requires understanding that cannot just be acquired by a short training course. It requires experience of both doing and observing. It is this experience that means they also fulfil the criteria below.

2. Have extensive knowledge (about pedagogy and knowledge for teaching).
It could be argued that if the purpose of an observation is just to collect data against pre-determined criteria, and it is not concerned with nuances of pedagogy or supporting teacher development, then maybe an observer who is not an expert teacher will suffice. But I would disagree. To start with, I think that maybe the key purpose of evaluation of teaching is to support teacher development. I also think that understanding the nuances of pedagogy is necessary. Useful observation requires subject matter knowledge and pedagogical content knowledge. In order to analyse a lesson or an episode in a lesson, an observer needs to understand what is happening in that lesson or that episode. For example, many observation tools refer to teachers using 'open questions'. However, I would argue that these are not easy to discern, even by many teachers, because the concept of open questions is far from simple. Whether questions are open or not is more to do with function than form. Identifying open questions is significantly more nuanced than just recognising whether questions begin with, 'Why…' or 'How…' etc. It requires observers to attempt to discern the purpose of the question and therefore try and understand the intention of the teacher in asking the question.

3. Are able to read and interpret (subtle) cues.
I have heard it suggested that non-specialists may be more impartial. In other words, teachers may not be impartial observers because they bring their own beliefs (and baggage) to an observation. But I would argue that apart from very low-inference observations, in which observers merely record whether something occurs or not, or count the number of times something occurs without making any judgement about quality, then observation is dependent on interpreting what is seen. Both experts and non-expert observers have to interpret, and therefore both have to learn to be impartial. Even non-expert observers bring their own beliefs into a classroom. However, experts have the advantage that their knowledge for teaching will help them to read and interpret what they see.

4. Notice what is significant.
In particular, for me, the key difference between expert and non-expert observers is that experts are more likely to notice what is significant. A number of researchers, including Grant et al (1998)[24] have commented on what teachers notice when they observe others teaching. In order to understand what is happening in a lesson, we first need to notice what is happening in that lesson, whether we are the teacher or an observer. It is particularly important that we notice what is significant in the lesson. The fact is, we all notice different things: we notice what we are sensitive to. A professional educator becomes sensitive to and notices particular things during lessons. It is obvious that the professional knowledge and experience of an observer affects what they notice; it requires experience and expertise to notice what is significant and to be able

[22] Coe, R. (2014). Classroom observation: it's harder than you think, CEMblog. Available at: http://www.cem.org/blog/414/

[23] https://k12education.gatesfoundation.org/blog/measures-of-effective-teaching-met-project/

[24] Grant, T. J., Hiebert, J. & Wearne, D. (1998). Observing and teaching reform-minded lessons: what do teachers see? Journal of Mathematics Teacher Education, 1, 217-36.

to understand why it is significant. Indeed, John Mason[25] identifies this as a characteristic of an expert:

> 'The mark of an expert is that they are sensitised to notice things which novices overlook. They have finer discernment.' (Mason, 2002, p.1)

I have many personal anecdotes that I could use to illustrate this. One example concerns a video of a mathematics lesson that was shown by a colleague (a non-mathematician) as part of a workshop that we were facilitating for inspectors. The lesson focused on a small class of primary children who were being taught the standard column method of addition involving two three-digit numbers. At one stage of the lesson, the class were asked to calculate 467 + 235 on their mini-whiteboards and then lift them up to show what they had done. A few children had clearly followed the algorithm correctly and got the correct answer. A few had written the numbers down but had not started the calculation, suggesting that they were unsure what to do next. One or two had started calculating by adding the digits from left to right (i.e. 4 + 2, then 6 + 3). My colleague commented to me that some of the children had obviously learned the method, but others hadn't. He remarked that one boy clearly didn't know what he was doing as there were about six lines of numbers on their whiteboard. As a mathematics specialist, I had noticed that this boy had used an expanded partitioning method and had got the correct answer, unlike many of his peers. I also noticed that the teacher didn't comment on what the boy had done. It would have been interesting to talk to the teacher about this boy, which brings me to my final attribute of expert observers, that they are able to provide useful feedback to teachers,

5. Are able to provide useful feedback to teachers.
Teacher development should be a key reason for evaluating teaching and observing teaching practice. But teacher development relies on teachers getting the most pertinent feedback on their practice, upon which they can reflect and act. This is reliant on the person observing the lesson having the knowledge and experience to notice what is significant, engage in dialogue about it, and help the teacher to identify the next steps in their development. I suggest that non-experts cannot give this level of useful feedback to teachers, and therefore cannot be as effective in supporting teacher growth. Unsurprisingly, evidence also suggests that teachers value feedback from experts (subject-specialists) more than from non-experts.[26] [27]

For these reasons, I argue that observations of teachers by 'non-experts' cannot be as valid or as useful as those carried out by 'experts'.

Discussion

A participant raised the question, 'What makes a person an expert?' She explained that she would probably be considered a mathematics expert by others, but in her own mind, she is an expert in an algebra but not in geometry. Expertise is therefore a question of perspective.

Another participant, who is a chemistry expert, explained that being an expert enables her to understand the logic of teaching chemistry. When observing a lesson, she sees what teachers are doing in that moment and how they are going to build on that, and she is also able to understand what has gone before and what the interconnections are, for example, with physics. Being an expert is about the 'connection knowledge' that can be made between concepts and ideas, as well as the 'sequencing knowledge'.

[25] Mason, J. (2002). Researching your own practice: the discipline of noticing. Abingdon: Routledge

[26] Evans, S., Jones, I. and Dawson, C. (2015). Do subject specialists produce more useful feedback than non-specialists when observing mathematics lessons? Proceedings of the 38th Conference of the International Group for the Psychology of Mathematics Education, 3, 33-40

[27] Peake, G. (2006). Observation of the Practice of Teaching. Huddersfield, UK: PCET Consortium, University of Huddersfield

Another participant agreed with the notion that observations should be carried out by experts, but asked,, 'What do we do if we don't have experts?' and 'How can we support non-experts (if that is what we have) to do something that is meaningful?'. A participant suggested that triangulating judgements using other sources of evidence beyond observation is a possible mitigation. For example, student voice is relevant. There are studies which point to the value of asking the students about their learning experience.[28]

Another participant wanted to ask similar questions about what can be done in situations where experts are unavailable, although he agreed that observation by experts being better was self-evident. He wondered whether we could move towards a clearer definition of what it means to be an expert. Furthermore, as part of that, is it possible to codify what a chemistry expert (for example) would see in a lesson, for the purpose of sharing this information with non-experts? Another participant explained that he had conducted research into the issue of inspections carried out by non-subject experts. The research found that the conclusions of non-experts were less valid than those of experts. The follow up question is therefore what should be done about this, as the answer cannot be to hire subject matter experts for every observation. The research suggested two answers to this question:

1. Codify the subject specific knowledge that is required to make sense of the subject as far as is possible. In the research project, subject expert groups were established to provide detailed codification of what it means to teach, for instance, history or mathematics.
2. The observers need to be thoroughly trained in using those codifications.

Another participant drew on her experience of running the TALIS Video Study. There were a considerable number of activities in the study where using a non-expert was acceptable. However, the video analysis could only be conducted by highly qualified experts. In this case, the research entirely depended on the expertise of those coding the lessons.

A participant raised the point that there is a difference between making judgements and observations which can be used as genuine professional development opportunities. Perhaps there is an argument that an expert is needed in order to make judgements about teaching, but if the observation is intended to be a learning opportunity for teachers, then generic questions which cause the teacher to reflect on their own practice do not necessarily need to be devised by a subject expert.

Another participant talked about the success she has had on her programme using non-experts strictly for collecting data. She said that they used engineering undergraduates because they tended to be precise and able to follow protocols. She found that sometimes the researchers found more issues using educational experts for observations because of their bias: this led the observers to focus on other things rather than what was in CLASS[29] (the observation tool they were using in the project). However, on the professional development side, she recognised the need for experts: where there is a component that involves coaching teachers on professional development, there is a higher level of credibility if teachers work with someone who has been in the classroom. Another participant supported this. She had found non-experts capable of data collection processes, including undergraduates of education and other disciplines. She said that they have also used the Quality Teaching Model with school-age students to give feedback to teachers, particularly in high schools, and found this to be effective. Non-experts or non-subject

[28] For example: Measures of Effective Teaching (MET) Project. (2012). Asking students about teaching. student perception surveys and their implementation. Bill and Melinda Gates Foundation. Available at: https://k12education.gatesfoundation.org/resource/asking-students-about-teaching-student-perception-surveys-and-their-implementation/

[29] https://curry.virginia.edu/classroom-assessment-scoring-system

experts tend to experience a lesson in much the same way as a student would. They ask questions about things that subject experts may sometimes neither question nor notice: for example, the experts take for granted why the teacher moved from one concept to another. The participant accepted that her work was mostly conducted in Australia, where they did not face the problems that others around the room had raised about teachers not having basic knowledge of their subject. She was concerned that if experts are relied on to say what good teaching is and if they are the ones to analyse practice, then teachers themselves may be disempowered from the process.

Another participant talked about a study she had undertaken which focused on early career mathematics teachers. The research methods included observations, for which specific protocols were developed. The findings of the study were that if 'context' encompasses both the subject and the content pedagogy, then it is necessary to have expert observers who are able to record whether mathematical mistakes are made. It is important to consider the purpose of observation. If it is accepted that subject knowledge competence is part of teaching, then subject matter specialists need to be part of any observation team. One participant raised the issue of whether there is capacity to do this. She questioned the feasibility of being able to find and use experts in different contexts.

Another participant requested clarification on the focus of the discussion, as some participants were speaking from a professional development perspective and others from a research perspective. Colin responded by saying that a key aspect of this challenge is to consider the purpose of any observation: naturally, some participants will look at it from a research perspective and some will look at it from a teacher development perspective, because those are the backgrounds people are coming from. There are many reasons for wanting to evaluate teaching quality, so a key question is: what is the purpose of doing so? Another participant agreed that this is a critical point. She felt it might be helpful to see if the group could agree on four broad purposes of classroom observation:

1. Teacher formative development;
2. Program impact evaluation;
3. Basic system diagnosis; and
4. Teacher accountability.

She elaborated on the fourth point (teacher accountability), stating that it is important to think about this, as a number of middle-income countries are struggling with how to move to a professionalised meritocratic teaching school system in which teachers are rewarded and incentivised on the basis of their performance. This requires a sophisticated and comprehensive type of evaluation; in this case, she agreed with Colin that the more expert the observers, the more valuable and valid those assessments are likely to be. However, she argued that for basic system diagnosis or for programme impact evaluation where a large number of observations need to be carried out, this would be very costly, so well-educated non-experts can be trained to use an observation tool effectively. Another participant strongly questioned the notion that classroom observations should be used for teacher accountability, due to the issues around validity, reliability and the need for triangulation. This participant argued that high-stakes observations are perhaps not so useful if system improvement is the goal. What is more important is professional development of teachers and the learning processes for initial teacher educators. Another participant agreed that it is not helpful to use either classroom observation or value-added measures to judge performance of individual teachers because the reliability and validity issues 'become serious'. Multiple observations of each teacher would have to be carried out to increase reliability and validity. Furthermore, this risked a whole set of perverse incentives being introduced into the system. In response, the participant who suggested observations as

part of high-stakes accountability questioned how does meritocratic system (in which teachers are promoted based on quality and not on years of service) get developed without the use of observations for accountability? She argued that while it is important that there is a suite of tools used to carry out evaluations, classroom observation should be one of the tools. This was refuted by the participant who said observations should not be used as part of high-stakes accountability decisions. When making decisions about promotions, much more additional information is needed. She argued that value-added data is useful for understanding processes that predict better student outcomes, but it is not good for judging teacher quality. It also encourages teaching to the test.

Another participant wanted to add to the list of purposes of observations: they can also be used simply to understand teaching, rather to evaluate impact or performance.

It was noted that there is another group of people in the [UK] education system who are doing a great deal of classroom observation: pre-service student teachers who have little to no expertise in observation There is an insistence that extensive observations form an early part of teacher training programmes, but students are not sufficiently trained to fully understand the quality of teaching they are observing. They are observing lessons in order to learn, rather than pass judgement on the teacher being observed. Another participant responded by saying that trainee teachers are being asked to learn how to notice things about teaching and things about learning. It is by doing this that beginner teachers get into a position where they enter the world of professional discourse, including discourse about their own subject. This participant also raised the point that there is a subtle difference between being an 'expert observer' and an 'expert who can observe'.

Another participant stated that the group needed to think beyond experts as only subject experts. There are also experts in how students learn and experts in how teachers support students to become increasingly independent. She stated that it is possible to support teachers to become expert observers of their own practice and each other's practices, particularly using a school-based or centre-based approaches.[30] There are some observation instruments that can be democratised if training is provided for teachers. This allows them to discuss and improve their practice. She suggested that teachers are capable and can become expert observers; furthermore, they have a vested interest in improving their own performance.

Colin closed the discussion by agreeing with the ideas expressed at the very beginning of the discussion: that the concept of expertise is challenging and that it is not absolute. He argued that it is important to continue returning to the question of purpose. He felt the discussion around when it is and is not appropriate to either have or not have experts is an important one. He thought the discussion on what can be done when there is a lack of experts available was critical, and it needs further discussion. He strongly agreed that we should view teachers in the system as potential experts. They could be useful resource when observers are needed. He spoke about how such experts could be used to develop and build system capacity, particularly in low-income contexts.

[30] For example, school-based teacher learning communities or communities of practice.

Challenge 4: Are 'snapshot' lesson evaluations as valid and useful as observations of full lessons?

Initial provocation: Ian Thompson, Associate Professor of English Education and Director of PGCE, University of Oxford

There is already a clear answer to this question: no, snapshots cannot be as valid and useful as observations of full lessons, just as observations of full lessons cannot be as valid as observations of a series of lessons.

However, it is important that we discuss this issue because it is the most common form of observation, particularly by schools themselves, and it will also be part of the new Ofsted framework for the inspection of schools [in the UK].

The major concern with snapshot approaches is that, depending on which moment you pick, you could either get a very good or very bad idea of the teaching; one that does not give an accurate overall picture of the quality of the pedagogy. It is also true that the snapshot approach is a source of much antagonism within schools. Teachers feel that this method puts pressure on them. Further, there is a question of judgements: how effective are the judgements and by what measures are teachers being judged?

All of these questions are valid and come into the overarching debate. However, there are five key points which are central to this issue.

The first issue is an ethical one: snapshot observations are both discourteous and unprofessional. If you are going to watch someone teach, then give them the credit of watching them teach properly. These kinds of snapshot observation are practically unheard of in other professions. Would we judge a play having watched only one scene?

Secondly, there is an issue of reliability and validity. All forms of observation are notoriously unreliable, either of evaluating teaching quality or student outcomes. Robert Coe has previously sparked debate when talking about this.[31] We can point to evidence from very well-resourced settings where observers are highly trained, or even accredited, but where the reliability of observations has been a big problem. Validity is also problematic, particularly in the case of peer observations, where teachers are often not given enough time to focus on preparing for observations. The 15-20 minute snapshot observations therefore become less likely to yield any sort of reliable or valid judgements.

Thirdly, snapshot observations can, at worst, become a one way process. Again, Robert Coe has previously pointed out that a problem with observations is that they tend to be very impressionistic. They are also prone to a degree of personal choice or emotional response. For example, if an observer is a proponent of group learning and the first thing they see during a lesson is group learning then s/he may be more predisposed to give positive feedback. However, the snapshot observation may not allow for the opportunity to understand how this group work is affecting the learning and where it sits in the relation to the rest of the lesson.

Fourth, snapshots can reduce observations to a focus just on student behaviours or classroom management. Experienced educators will know that it is sometimes during the most uncomfortable moments in a lesson, where the learning actually happens. However, it is very difficult for a teacher who is being observed to allow moments like that to happen.

[31] For example: Coe, R. (2014). Classroom observation: it's harder than you think, CEMblog. Available at: http://www.cem.org/blog/414/

Finally, observation may lead to teachers to using safe or conservative teaching methods. Teachers are sometimes afraid to be caught out and, in some cases, will even use pedagogies that they know to be wrong because they know they will be expected by the observer.

In conclusion, while snapshots are not as effective as observations of full lessons, where snapshots are used, they are not necessarily being used to their full potential. This could be rectified in order to improve the value of these kinds of observation.

Discussion

A participant asked whether there is any merit in very brief engagements in lessons. He talked about a group of headteachers he was working with in Vietnam who were asked to keep diaries. The headteachers spent around 60% of their time walking around school, seeing what was happening, engaging with other professionals. There was a culture where people were in and out of classrooms: there may be insights that are drawn from such a culture which are deserving of merit. Ian replied that there is merit, but not where these snapshots are used for accountability or measuring performance. Another participant agreed that 'learning walks' are not bad *per se*, but they could be if conducted for high-stakes purposes. This participant also stated that 20 minutes is simply not long enough to see how students come into a class, how learning is introduced and related to previous learning, if the students have been completing homework and if the teacher sums up learning at the end. The participant felt it was better to look at the full lesson, even if that meant looking at fewer lessons overall.

Another participant asked about using this technique in a low-income setting where time is limited: would it be preferable to have one full observation rather than three shorter observations? Ian argued that one full lesson observation would be better. Another participant said that snapshots may have a purpose and be of value if the culture is right and there are high quality conversations taking place after the observation. Another participant talked about a study she had been involved in where they used the 'Teach'[32] observation tool from the World Bank. This is based on ten-minute snapshot observations in a lesson [followed by ten minutes recording]. This model creates times in the lesson where no observation is taking place while the observer is recording, so things happening in that part of the lesson may be missed. To overcome this, the researchers developed some additional items which allowed them to observe in a more holistic manner along with some mathematics specific observation prompts. She stated that there was both high reliability and high validity in the study.

Another participant was puzzled by the challenge to snapshot observations because she had personally carried out a great deal of work in Latin America using the Stallings[33] instrument, whereby the entire class is observed to try to ascertain basic parameters about the teacher's use of time, materials and pedagogical techniques. Multiple teachers in each school are observed to get a sense of the variation in practice in the school. Having said that, she expressed frustrations with the limitations of the instrument and the fact that there are parts of the lesson which are not observed. This could be time used to gather supporting qualitative data about the teacher's practice. She is currently working with a team to merge Stallings with the World Bank's 'Teach' tool (which combines precise, validated quantitative data with qualitative assessment of quality of instruction, support for socio-economic development and classroom culture). Another participant asked how different Teach was to Stallings, since the observation is performed in snapshots and there is a time on task element. The participant who is merging the two instruments stated that Teach's time on task element is very limited.

[32] http://saber.worldbank.org/index.cfm?indx=5&sub=7

[33] https://www.worldbank.org/en/programs/sief-trust-fund/brief/the-stallings-classroom-snapshot

One participant referenced the Measures of Teacher Effectiveness (MET) project[34] which concluded that shorter lesson observations were better value for money in terms of their reliability. Another participant said that it depends what it is that one wants to measure. If the purpose of the observation is to find out time on task, then it is important that a whole lesson is observed. If the purpose of the observation is to find out about the qualitative aspects of teacher-student interaction, it may be more feasible to do that within a segment of a lesson. Another participant noted that even with this, the limitation of a 20-minute segment is that there will be things than can be missed (for example, whether the teacher has worked with all student groups).

[34] https://k12education.gatesfoundation.org/blog/measures-of-effective-teaching-met-project/

Challenge 5: Is it realistic for evaluation of teaching to be a teacher developmental process?

Initial provocation: Trevor Mutton, Associate Professor, Director of Professional Programmes and Director of the Oxford Education Deanery, University of Oxford

There seem to be three possible responses to this question. First, 'Yes, of course'; second 'No, probably not'; and third 'It depends'.

The case 'against':

We must ask what the purpose of evaluating the quality of teaching actually is. In whose interests is the evaluation being carried out?

If evaluation is a process of making value judgements about a level of performance or achievement, then it is inherently summative and is not, therefore, the best way to inform the process of professional development. Evaluation requires measurement, and an instrument for carrying out that measurement – usually evaluating performance against external criteria or often some sort of standard. If evaluation is the process of observing and measuring teaching for the purpose of judging it, by comparing it to a standard, then it can only lead to development within the context of a system of performance management review. So, a fundamental part of such a process is measuring teacher performance as a mechanism for continually improving skills and outcomes, perhaps not as part of a wider teacher development process. This all comes down to how the latter is seen within a broader view of teacher professionalism. Julie Evetts[35] has talked about the way in which increased accountability has led to an attempt to 'measure and demonstrate professionalism' (Evetts, 2011, p.412), which in turn reminds us of the tensions between what Judyth Sachs (2010)[36] has characterised as 'managerial professionalism', which is driven by an overriding emphasis on accountability and effectiveness, and 'democratic professionalism', which emerges from the profession itself (and thus acknowledges the scope for judgment in complex situations, underpinned by expert knowledge).

The Organisation for Economic Co-operation and Development (OECD), reporting in 2013[37] on the most recent TALIS survey, did identify the need to align performance management with professional development opportunities:

> 'Without a clear link to professional development opportunities, the impact of teacher appraisal and performance review will be relatively limited.'
> (OECD, 2013b, p.62)

However, it goes on to identify just how it sees this process working, when it says:

> 'Identifying individual teachers' strengths and weaknesses helps to determine which professional-development activities meet the teacher's own needs as well as the school's priorities.' (OECD, 2013b, p.66)

This is a top-down process which is about identifying strengths and weaknesses which are then to be addressed through a management process – by determining subsequent training activities which will, presumably, be put into place to deal with any deficit. Such a process

[35] Evetts, J. (2011). A new professionalism? Challenges and opportunities. Current Sociology, 59(4), 406-422.

[36] Sachs, J. (2010) Teacher professional identity: competing discourses, competing outcomes, Journal of Education Policy, 16 (2), 149-161

[37] OECD (2013). Teachers for the 21st Century: Using evaluation to improve teaching. Paris: OECD

only leads to professional development if one conceptualises such development in terms of a process of addressing targets identified by others. As Kerry Elliott[38] from Melbourne argues:

> 'A process that is seen as a means to 'manage' teachers needs to be reconsidered if a credible performance appraisal system is to be accepted by them (Ingvarson, 2012).' (Elliott, 2015, p.110)

So, could such a re-consideration support an alternative?

Let us now consider the case 'for':

The first thing to say is that the relationship between the evaluation of teaching and the potential for subsequent professional development is undoubtedly complex. That is not to say, however, that the evaluation process cannot be implemented in such a way as to facilitate ongoing professional learning. Helen Timperley and her colleagues[39] found that:

> 'feedback from observations assisted teachers to translate theoretical principles into practice.' (Timperley et al., 2007, p.xxxvi)

However, they also found that:

> 'while observation and feedback may support teachers to implement or refine new practices in ways that have a positive impact on students, they do not guarantee it.' (Timperley et al., 2007, p.86)

Kerry Elliott (2015) identified that one of the factors involved in professional development leading to improvements in student achievement was when:

> 'Approaches were responsive to learning processes. Engaging teachers in the process and challenging their existing ideas and assumptions was important in developing congruence between new information and practice.'
> (Elliott, 2015, p.108)

Teachers, therefore, have to be part of the process, not the people to whom the process is applied. So is it possible to imagine a way in which evaluation might lead to meaningful professional learning? This would, of course, entail a very different approach. This leads to some questions.

First, what does the observer/evaluator need to know and to understand before the teaching is observed – in particular, with respect to what the teacher is trying to achieve. Context is vital in this respect. Second, who sets the criteria against which the teaching is to be evaluated? Supposing it is the teacher being observed who selects the evaluation criteria, would it matter if these evaluation criteria differed from teacher to teacher? Or it might be that these are criteria that have been commonly agreed across a group of teachers within a school, or more widely. Third, what evidence would need to be collected and how would this happen? Does the observer then become the data collector rather than the evaluator? Fourth, how is the judgement made? Could this be a judgement reached jointly and collectively by the teacher being observed and the observer, through a detailed discussion following the observation itself, in which the evidence is examined and conclusions drawn? Fifth, who then decides on an appropriate response, by way of subsequent professional learning? Perhaps, again, this would be the teacher in collaboration with the observer.

[38] Elliott, K. (2015). Teacher Performance Appraisal: more about performance or development? Australian Journal of teacher education, 40(9), 102-116

[39] Timperley, H., Wilson, A., Barrar, H., and Fung, I. (2007). Teacher professional learning and development: best evidence synthesis iteration. Wellington, New Zealand: Ministry of Education

This approach is the one which we try to encourage within our teacher education programme, particularly in the latter stages of the programme when the student teachers have reached a stage where we consider them able to take responsibility for their ongoing professional learning.

Discussion

A participant wanted to reflect upon the democratisation of observation tools with those who perhaps do not work in the English context. The pushback may be that some teachers are not ready for this approach. Another participant noted that in many of these settings, there is little to no data available on teacher evaluation, so any tool that gives some information could be productive in helping countries and governments. Another participant reiterated the point about the importance of context. For example, if in Mozambique there are 60% of teachers not in school, the starting point will be less around teacher development and more around incentives that encourage teachers to attend school. There is also a question of the level of authority which is available to a headteacher to incentivise or sanction teachers to improve performance. If there is an absence of real power, a headteacher does not have the levers at his/her disposal to drive performance. Another participant agreed that involving teachers themselves in the process is important so that they critically reflect on their own practice.

Another participant wanted to share her experience of using CLASS[40] in Latin America and South East Asia. Most people recognise CLASS as an evaluation tool, but there are also successful professional development programmes based on CLASS. It is a partially democratic model and is strength-based professional development for teachers, in which coaches with the evaluation data do cyclical evaluations. Within the process, teachers have some flexibility in choosing which areas to focus on. Another participant raised the idea that there can be a disconnect between observation tools such as CLASS being used in countries which may have their own set of standards. There can therefore be confusion around what is being promoted as good practice or what makes someone a competent teacher in a particular context. Another participant said that there is a question over efficiency and paying attention to the concerns of teacher. For example, when working with beginner teachers, the observer will use the standards but also talk to the teacher. If the teacher is aware of some of the issues, then it might be wise to focus on those. Another participant talked about teacher self-evaluation and how this might be used as part of evaluation.

One participant talked about a particular tool that he had used. The key point was that the tool was transparent and accessible to the teachers. This allowed it to become embedded into everyday practice. What was particularly effective teachers could talk to each other and review their practice together using the tool, deciding the most appropriate descriptions for where they each were and where they needed to go.

Trevor wrapped up the discussion by re-emphasising the importance of purpose. Beyond this, it is important to think about what particular instruments can and cannot do. There are also some questions about what happens when a framework is taken to scale, as the level of flexibility in the framework will then not be as great.

[40] https://curry.virginia.edu/classroom-assessment-scoring-system

Challenge 6: Can we evaluate teaching quality fairly without taking into account the context and previous experiences that teachers have had, including their knowledge and beliefs?

Initial provocation: Teresa Tatto, Southwest Borderlands Professor of Comparative Education and Professor of Educational Leadership & Innovation, Arizona State University

The question posed by challenge six has many elements to it and so answering it could take us down several different avenues. This presentation focusses on one particular aspect: what will it mean to develop a programme of research for the profession and by the profession?

Considering definitions

Even in answering this question, there is a considerable amount of interpretation involved. Several key words require definition and we must also consider the challenges of evidence, challenges of usefulness, and challenges in engaging and collaborating with others, particularly teachers and teacher educators.

To begin, it is important to ask ourselves what we mean when we use certain key terms. First, what do we mean by teacher quality? This is something that is different depending on the context where teachers teach, and it is often difficult for researchers within and across countries to agree on a working definition. An approach that can be useful here is to think about teacher quality as something aspirational. For instance, what do we want teachers to know and be able to do once they have completed their initial training, and after their first years on the job?

Second, we must also consider what we mean by evaluation. This also comes back to how and why we are evaluating in the first place, because the answer differs depending on the purpose of the evaluation, on what are we attempting to measure and why, and on how we are planning to use the results. Evaluations of teacher quality (as currently implemented) provide very limited information and may lack validity as actual indicators of quality, such as evaluations done using value-added models which assess teacher effectiveness by aggregating differences between the predicted performances of a teacher's pupils and their actual performance after some period of instruction (Harris, 2011)[41]. Other types of evaluation, such as a needs assessment approach, may be more promising, as they can help identify the elements that may be missing as we attempt to accomplish specific goals (see, for instance, Tatto and Pippin, 2017).[42]

Third, evaluations must be fair. When we use the term 'fairness', we must consider the learning opportunities that teachers have had when learning to teach. For instance, it would be very difficult to demonstrate high-quality teaching if an individual has not had the opportunity to learn the essential features that can be used to define teaching quality, such as the knowledge of the subject they are expected to teach, the pedagogy of the subject, and the creation of a conducive learning environment for all pupils, among others.

---

[41] Harris, D. N. (2011). Value-added measures in education: What every educator should know. Cambridge, MA: Harvard Education Press.

[42] Tatto, M. T. and Pippin, J. (2017). The quest for quality and the rise of accountability systems in teacher education. In J. D. Clandinin and J. Husu (Eds.), International Handbook of Research in Teacher Education. Los Angeles, California: SAGE

Considering measurement

Considering what we mean by the evaluation of teaching quality is important because the concepts that are chosen to define teaching quality will have consequences for measurement. Thus, a key question is: what elements should be included in such a definition, and how can we justify their inclusion? Of course, such justification should be made according to research evidence, plus other evidence relevant to the contexts and cultures in which teachers practice. Here, I argue that to measure teaching quality fairly and productively, the following elements, at a minimum, would need to be included:

- the preparation teachers have received (expressed as opportunities to learn), and the quality of the programmes in which teachers have been trained;
- the knowledge that teachers possess on the subjects they are asked to teach, and of other aspects that teachers need to be effective (e.g., pedagogy, assessments, cognition, research, communities);
- the beliefs that teachers hold about what it takes to learn different subjects; and,
- the differences in the ability of their pupils, and of the contexts in which they teach.

If we do not take these and other similar aspects into consideration, then we are not accounting for important elements that are determining factors for teacher quality and we are therefore ignoring vast amounts of contextual and explanatory data. When we consider large scale comparative studies, this becomes an imperative and a highly complex endeavour.

Considering evidence

We must also consider the challenge of evidence. There is a lack of valid and reliable evidence for all of the different reforms that have been introduced in different contexts and countries, including the evaluation of teaching quality – especially when we consider that most studies do not define key terms and measurements in advance or work to understand the nuances of each context.

Attempts at evaluating teaching quality have long been undertaken, , without significant evidence of its impact on teaching and learning, particularly in under-resourced settings. The question here is: what evidence should we gather to better explore whether mandates such as the evaluation of teaching quality have contributed to improvements in teaching, teacher education and, ultimately, on pupil learning and retention?

Clearly, it is important to evaluate teaching quality. However, it needs to be done for an educational purpose and that educational purpose must be learning. To date, it is not clear whether evaluations of teaching quality have resulted in significant learning outcomes for either teachers or pupils – or, for that matter, for teacher educators. This is true not just in low-resource settings but also in the U.S., where the 'No Child Left Behind' Act of 2001 created an 'outcome frenzy' with no palpable benefits for teachers or their pupils.[43]

[43] In 2001 the *No Child Left Behind Act of 2001* (NCLB) was passed (U.S. Congress, 2002). This Act of Congress which re-authorized the *Elementary and Secondary Education Act of 1965* (U.S. Congress, 1965) introduced a system-wide framework for standards-based education reform including measurable goals to improve individual outcomes in education. The standards came accompanied by the introduction (by the states) of a more formal system of school and teacher evaluations. This was followed by calls from multiple fronts for close scrutiny of traditional teacher education programmes' curricula and outcomes, as poor results in teacher evaluations were seen as a result of the 'mediocre' preparation teachers received in their programmes. To date, it is not clear whether the introduction of standards, testing and evaluations of teaching quality have improved learning for teachers or their pupils. (Tatto, Burn, Menter, Mutton and Thompson, 2018, p.64-65)
Tatto, M. T., Burn, K., Menter, I., Mutton, T., and Thompson, I. (2018). Learning to teach in England and the United States: The evolution of policy and practice. Abingdon, England: Routledge.

Considering usefulness

It could be argued that evaluations of teaching quality are not always useful. It really depends on factors such as the purpose of the evaluation, who conducts the evaluation and how, and what is done with the results.

There is a case for involving teachers and teacher educators in studies of their own practice and in the production of useful knowledge. While the outcomes of these studies could be used to respond to accountability demands, that would not be their main purpose. The main purpose of evaluations of teaching by educationalists should be to use these as a building block to develop a programme of research for the profession and by the profession, to learn from and to improve practice in both teaching and teacher education on a recurrent and periodic basis. Developing such a programme of research is necessary, yet such an effort, if taken seriously, requires time and resources. Our experience indicates that it can take up to a year to arrive at agreed definitions and measurements, particularly in the case of cross-country comparative studies.

Further, given the underdeveloped capacity that afflicts the field, it may take time to develop the capacity of teachers and teacher educators to do research on their own teaching and teacher education programmes. This is an ambitious yet necessary agenda for the profession that depends on obtaining needed resources and building local capacity.

The development of a research agenda that is truly useful for the profession is not without challenges. Indeed, while local efforts by educationists are emerging (and have emerged already in a number of countries), international large-scale data collection exercises using models from economics threaten to override emergent local research, in part due to confusion about what these expensive endeavours can and cannot do. The Organisation for Economic Co-operation and Development (OECD) surveys serve as the best example of this trend. Country governments invest significant resources in these large-scale collection exercises while limiting investment in local research. This is a misguided strategy. While these data collection exercises may offer to provide policymakers with a possible argument to push policies (often those prescribed by the OECD), in practice, these data collection exercises only provide at best indicators of a construct for a very limited proportion of the population in a discrete moment in time (e.g. via the PISA test). The data obtained from PISA cannot be used to explain, for instance, why test results go up or down or how to address underlying curriculum and similar issues; and they certainly cannot be used to improve practice. For that, a more holistic, local exploration is needed. To be useful, evaluations of teacher quality need to be conducted by educationists, need to be local (i.e. at the individual teacher level) and holistic (i.e. encompassing education systems at the macro, meso and micro levels) to uncover why and how things change and with what consequences. Local efforts are not only more likely to be useful, but are also more likely to be sustainable.

Considering agentic collaboration

Evaluations must be conducted in a way that involves those who are 'being evaluated' – in this case, teachers (and teacher educators) – as agents. Evaluators must take the time to understand specific aspects of teachers' experience, learning opportunities, and contexts, for such work to contribute to the development of the teaching profession. Currently, in a number of countries, evaluations of teaching quality and teacher education are rare or non-existent, a situation that has allowed non-educationists to attempt to fill the gap and to begin to control what information is collected and how it is to be used. As already mentioned, the most notable example is the OECD, with its indicator collection exercises and more, recently, focusing on teaching and teacher education (such as TALIS, Flying Start and Teacher Ready!, and a proposed study of teacher educator's knowledge and pedagogy).[44] In what amounts to an all-encompassing

[44] http://www.oecdteacherready.org; http://www.oecd.org/education/school/talis-initial-teacher-preparation-study.htm; https://read.oecd-ilibrary.org/education/a-flying-start_cf74e549-en#page1

strategy working with in-country Ministries and Departments of Education, the OECD has launched its Education 2030 Project, which has already developed the 'OECD Learning Framework' to 'define a clearer vision and goals for the future of education systems'.[45]

In contrast, and as a necessary alternative to the exclusive logic of economic growth, we find important international comparative studies which have demonstrated how methodologies can be developed collaboratively to be used locally to produce research that supports the teaching profession (Tatto et al., 2013[46]; Tatto et al., 2018[47]; Tatto et al., forthcoming[48]). These studies by teacher educators and teachers were conducted to first understand the processes and outcomes of teacher preparation programmes, as well as the curriculum standards of programmes and schools for all countries that participated. The studies developed instruments with validity evidence to measure the (aspirational) outcomes of teacher education, including assessments of content knowledge, pedagogical content knowledge, and pedagogy, as well as questionnaires that helped us to understand teachers' learning opportunities and beliefs. The research team then worked with new teachers to help them learn how to research and evaluate their own practice via the use of interviews, observations and evaluations of their pupils' learning. These studies successfully connect teacher education with teaching quality. With interdisciplinary teams of educationalists, teacher educators and teachers, these studies have developed a unique methodology and have designed valid and reliable tools that can be used for the self-study of teacher preparation and teacher quality among early career primary and secondary school teachers on a regular basis. These are important examples of studies that have successfully included teacher educators and teachers in studying their own practice. Thus, an important question for educationalists is, now that we have applicable self-study models, what would it take to regain agency to develop the profession in the same way that other professions have evolved in order to self-regulate, evaluate and improve our own practices?

The argument is not whether we should evaluate teaching quality (and teacher preparation quality) but rather how to use these efforts to build capacity to engage in fair, productive and sustainable evaluation systems for and by the teaching profession.

Discussion

One participant noted that the discussion had almost come full circle. The two main areas of discussion had been around context and purpose. The participant picked up on the notion that evaluating teaching quality had been happening for a long time without evidence of impact. They questioned whether it was really no evidence of impact, or whether there was simply no evidence available Teresa said that many of the studies on teacher effectiveness had been carried out in the late 1960s and 1970s. They were carried out in the most part by economists from the World Bank. The contribution was helpful at the time because they were the first ones who were talking about teachers' importance, the importance of teacher learning and the importance of teacher quality. However, there were no real efforts made to define the essential characteristics of teaching and of teachers. For example, knowledge was measured by indicators such as number of years of study, whether teachers had certain credentials, or the number of years of experience. What teachers knew in terms of content and pedagogy was not really measured. Even now, looking at studies published in the USA after 'No Child Left Behind' (using the data

[45] http://www.oecd.org/education/ceri/strategic-education-governance.htm
[46] Tatto. M. T. (ed.) (2013). The Teacher Education and Development Study in Mathematics (TEDS-M). Policy, Practice, and Readiness to Teach Primary and Secondary Mathematics in 17 Countries: Technical Report. Amsterdam: International Association for the Evaluation of Student Achievement
[47] Tatto, M. T., Rodriguez, M., Smith, W., Reckase, M., and Bankov, K. (Eds.) (2018). Exploring the Mathematics Education of Teachers using TEDS-M Data. Dordrecht, Netherlands: Springer.
[48] Tatto, M. T., Rodriguez, M., Reckase, M., Smith, W., and Pippin, J. (forthcoming). The First Five Years of Teaching Mathematics (FIRSTMATH): Concepts, Methods and Strategies for Comparative International Research. Dordrecht, Netherlands: Springer

collected), the idea of teacher knowledge is not really included in the questions. Teresa stated that in the research that she and colleagues had carried out, they attempted to do this, and in doing so, they uncovered the enormous amount of disagreement that there is in the field. There is also the point that educators themselves – as opposed to economists –are designing and carrying out the studies, there is a better chance of really finding out what is going on.

Another participant felt that there had been some studies that had provided evidence evaluations. She argued that whilst there were not large numbers of studies, she disagreed that there was no evidence of impact. She conceded that these were normally small-scale and on specific aspects, for example, of behaviour. Teresa accepted this but said that studies which have validity and reliability on multiple aspects of what it means to be a teacher (particularly the relationship between teacher knowledge and practice) need to be longitudinal, and these do not really exist. Another participant noted that while there is some evidence in low- to middle-income counties on evaluating teaching through observation, there is not a lot. There is little to nothing on teacher knowledge. Another thing that surprised this participant is that given that there is acceptance that teachers' beliefs influence their practice, there is very little evidence on this, apart from small-scale research on self-efficacy. Teresa said that they had found a strong correlation between teachers' beliefs about mathematics and the quality of teaching. They found that teachers who believed that mathematics could be taught as a series of rules and procedures had low scores in terms of both mathematics subject knowledge and pedagogical content knowledge assessments. This primarily stemmed from the way in which these teachers themselves were taught mathematics. One solution would therefore be to teach them enquiry-based mathematics pedagogy. Another participant queried the term 'fairly' when considering the statement that we need to understand teachers' knowledge and beliefs in order to evaluate them fairly. He felt that a sense of unfairness emanates from the point of view of the teachers and teaching profession. From the perspective of a student, it does not make any difference. This idea was challenged by Teresa: she argued that the way a person has been taught tends to be reproduced if that person becomes a teacher, and therefore it does matter to the student. Another participant agreed, saying that it is important not to forget that teachers' views must often be challenged. There are a number of problematic beliefs held by teachers about how students learn, or how certain subsets of students learn.

Other issues that participants felt had not been covered by the challenges or discussions

When asked to raise any other issues related to evaluating teaching quality that they felt had not been sufficiently addressed during the symposium, participants identified the following:

1. We did not pay enough attention to the students' experiences and student voice.
2. We need to spend more time discussing what all this means in practice, particularly when we are not in an 'ideal' context.
3. We did not talk at all about the challenge for developing countries to move from automatic promotions, low standards and low professionalism towards a professionalised, meritocratic teaching force, particularly when there is limited capacity to do so. The example of Mexico, whose teacher evaluation reform programme has collapsed, demonstrates just how difficult this is for policymakers.
4. We did not talk enough about the potential for developing longitudinal studies that can be implemented by teacher educators but feed through the first few years of teaching. Such studies would empower teachers and teacher educators, building the profession from the ground up, rather than investing money in sporadic large-scale studies.
5. We need to consider how we engage teachers themselves, so that we make sure we do not advocate something that disengages the teachers.
6. It would be helpful to discuss and explore the necessary capabilities of the system that need to be in place before we can begin to implement some of the approaches that we talked about.
7. We did not discuss the sharp drop in the number of people who want to become teachers and the large numbers who are leaving the profession.
8. We did not discuss the importance of teachers' beliefs and motivation, and approaches to evaluating them.

Moving forward

At the close of the symposium, participants made a range of recommendations for next steps and future work. The organisers and participants agreed they would identify and reach out further stakeholders – notably including teachers and policymakers – to disseminate the conversation further and ensure that these new stakeholders were included in future discussion. The OECD, the Global Partnership for Education (GPE) and Research on Improving Systems of Education (RISE), along with the wider research community, would be helpful contacts in developing this work in future. Further discussions and research will be necessary, specifically on the question of purpose in teaching quality evaluation. Moreover, an important focus of future research will be on the practical approaches to evaluating teaching quality in low- and middle-income countries.

There are clearly many important issues that will require further thought, reflection and analysis moving forwards. We look forward to exploring these issues in future research and discussion.