Performance over professional learning and the complexity puzzle: lesson observation in England's Further Education sector

Matt O'Leary[1] (a) and Phil Wood (b)
Centre for the Study of Practice and Culture in Education (CSPACE), Birmingham City University, Birmingham, UK (a); School of Education, University of Leicester, UK (b).

Abstract
Attempts to measure the quality of teaching and learning have resulted in an overreliance on quantitative performance data and the normalisation of a set of reductionist practices in England's Further Education (FE) sector in recent years. Focusing on lesson observation as an illustrative example and drawing on data from a national study, this paper examines the application of observation and its impact on FE teachers' practice. In viewing lesson observation through a complexity theory lens and contextualising it against the wider neoliberal backdrop of the marketisation of education, we seek to critique the inadequacies of current reductionist approaches to teacher evaluation, whilst simultaneously opening up a debate regarding the consequences of seeing classrooms as complex adaptive systems. In focusing on performative models of lesson observation in particular, this paper exposes what we perceive as some of the epistemological and methodological shortcomings of neoliberalism in practice, but also offers an alternative way forward in dealing with the contested practice of evaluating the quality of teaching and learning.

Keywords: lesson observation, teacher evaluation, complexity theory, neoliberalism, marketization, professional development

Introduction

This paper argues that graded lesson observations have come to embody a marketised, atomistic approach to capturing the complexity of teaching and learning. We maintain that such performative models of observation seek to reduce the complex processes of classroom interaction to a superficial set of skills and behaviours in an attempt to quantify what is debatably unquantifiable in a consistently reliable way.

In this paper we explore the practice of lesson observation through two distinctive lenses. The first sets the topic against the wider neoliberal backdrop of the marketization of education and the reliance on metrics that seek to quantify and evaluate the quality of teachers' work. The second makes use of aspects of complexity theory to examine the use of lesson observation. The findings and discussion section of the paper draw on research

[1] Corresponding author: matthew.o'leary@bcu.ac.uk

data from a national project carried out recently in the Further Education (FE) sector in England. The data discussed reveal some of the counterproductive consequences of graded observation on the professional lives of staff working in FE, whilst also exposing how such reductive approaches to observation fail to see the complexity of the pedagogic processes that exist in educational settings.

The FE sector in England has traditionally been located between secondary schools and universities. While there are similarities in the curriculum offered in FE and schools with both providing education for teenage students, there are equally some fundamental differences between the two. FE offers a wide array of vocational subjects, work-based learning and community provision. Unlike schools, FE also caters for a large population of adult returners to learning, typically those seeking to improve their qualifications and/or gain new skills later in life. It is little surprise therefore that FE is frequently referred to as a 'diverse' and 'complex' sector (Huddleston and Unwin 2013). This is also reflected in the size of some organisations i.e. ranging from large FE colleges catering for over 10,000 full-time and part-time students to small training providers with less than 50 students on the register.

The final point to make here about FE is that it has a history as one of the most market-tested areas of public sector reform in England with successive governments opening up its provision to the forces of marketization and the accompanying technologies of managerialism and performativity (O'Leary 2015). It is this feature of FE's recent history, in particular, that has had the biggest impact on the way in which observation has evolved and been used as a tool of teacher evaluation over the last two decades. But the policy backdrop in which this project took place is also worth highlighting briefly at the outset.

Lesson observation is a contentious issue that has provoked a lot of discussion in FE over the last decade. In some cases the level of discontent experienced by staff has led to industrial disputes, including the boycotting of observations. In those situations where industrial disputes have occurred, this has often been in response to what practitioners have perceived as punitive observation policies imposed on them by senior managers. One of the main aims underpinning the use of graded observations is the identification and separation of effective from ineffective teachers and, in some cases, eventual dismissal for those

characterised as ineffective. Such policies have ostensibly linked the outcomes of graded observations to disciplinary or capability procedures. It is therefore no exaggeration to say that the backdrop to this research project was one in which observation had become an emotive and disputed area of practice for those working in FE.

The first half of this paper begins by situating the emergence of performative models of observation against the wider neoliberal backdrop of the marketization of education and ongoing attempts to measure the quality of teaching and teachers both nationally and internationally. In drawing on aspects of complexity theory, the discussion is then opened up to consider classrooms as complex adaptive systems and what the repercussions might be for how observation is conceptualised and used in the context of teacher evaluation.

The second half of the paper presents an overview of the project's research design in which the research focus, sample and research methods are outlined. This is followed by discussion of the project's findings, focusing on two key themes from the research data that capture the central argument put forward in this paper relating to: 1) the counterproductive effects of performative models of graded observations on FE teachers' practice and 2) how such metrics-based mechanisms like graded observations fail to capture the complexity of classrooms and pedagogy as they are, by design and purpose, overly simplistic and reductionist.

The paper concludes with a discussion about what the findings imply for the future use of lesson observation in an educational context and recommendations on how to move forward.

The importance of capturing the quality of teaching and teachers

Since the advent of neoliberal education policy and the proliferation of new public management (NPM) in the English education system in the 1980s, there have been ongoing attempts to measure teacher performance. In a system that has become driven by the demands of audit and accountability (Power 1994), lesson observation has become an ever-present mechanism in the professional lives of teachers, used largely as a tool for evaluating their classroom performance. Arguably what has led to this reliance on observation as a

source of evidence of teacher quality is its convenience as one of many reductive tools used to drive a neoliberal system, what Ball (2012, 30) refers to as 'a set of moral technologies' (e.g. quality assurance, target setting, continuous improvement).

The most recognised of these tools is the 4-point grading scale used by the inspectorate Ofsted, the body responsible for monitoring and assessing educational provision in England. The 4-point scale is the epitome of a tool designed to satisfy the neoliberal obsession of trying to quantify and measure all forms of human activity, enshrined in the oft-quoted saying 'you can't manage what you can't measure', a maxim that has twin roots in reductionism and a marketised approach to educational improvement, which attempts to make the complex simple and measureable. In this instance, graded observations seek to transpose reductionist and standardised forms of measurement to the complex and unpredictable phenomena which together form human behaviour and interactions. As Morin (2008, 39) argues:

> The paradigm of simplicity puts order in the universe and chases out disorder. Order is reduced to one law, one principle. Simplicity can see either the one or the many, but it can't see that the One is perhaps at the same time Many. The principle of simplicity either separates that which is linked (disjunction) or unifies that which is diverse (reduction).

Models of graded observations have become normalised as a key mechanism for assessing and monitoring the quality of teaching and learning by employers and external agencies alike since the 1990s (O'Leary 2013). They epitomise the economisation and marketisation of education. These 'graded observations' are purportedly summative assessments of a teacher's classroom performance and overall competence, culminating in the award of a grade based on Ofsted's 4-point scale used as a central measure in inspection frameworks. It is only very recently that Ofsted announced a policy shift away from the practice of grading individual lesson observations in FE inspections (Morrison 2015). The extent to which this new policy has been embraced by senior managers across the sector to date is patchy at best, suggesting that the grading of teacher

performance is a deeply engrained practice with some refusing to move away from it (e.g. Exley 2014).

Global interest in improving education systems has risen sharply in recent years. Fuelled by the ever-growing importance of comparative performance data from international assessment systems such as the OECD's flagship Programme for International Student Assessment (PISA), the drive for continuous improvement in educational standards has undoubtedly become a high priority for many governments worldwide (Meyer & Benavot 2013). With increasing links being made between a country's economic competitiveness and its levels of educational achievement, politicians and policy makers are keen to identify a recognised formula for success. The importance of teachers and the quality of teaching in student achievement have figured prominently in recent studies, with particular interest in research exploring teacher effectiveness in the hope of pinpointing the skills and qualities displayed by the 'effective teacher' (e.g. Darling-Hammond 2005; Stronge et al 2011).

Darling-Hammond maintains that current models of teacher evaluation for accountability purposes fail to provide reliable information about teachers' professional competence and performance in the classroom and that they do little to enhance teacher learning. Rather than focusing efforts on increasing the reliability of the methods of evaluation, Darling-Hammond (2014, 5) recommends that any effective system of teacher evaluation needs to be conceptualised 'as part of a teaching and learning system that supports continuous improvement, both for individual teachers and for the profession as a whole'. She sees teacher evaluation as a facet of a holistic system of teacher learning and continued professional growth rather than as a disconnected monitoring exercise carried out by the senior managers of an institution. She goes on to argue that the development of collaborative communities of teacher learning 'will do more to support student achievement than dozens of the most elaborate ranking schemes ever could' (ibid), thus emphasising the importance of teacher growth as a collective act fostered through collegial communities of practice rather than individualistic competition.

Observation remains a key process by which teachers' effectiveness is measured in England. Hence, there has been a lot of discussion recently amongst policy makers and practitioners

over the use of lesson observation as a method of assessing the quality of teaching and learning (e.g. O'Leary 2014). In Ofsted, much of this discussion has converged around how observation is used as a source of evidence during inspections and particularly the issue of grading individual lesson observations. Rising criticism from informal teacher groups, centring largely on the inability to see much of the learning process through observation (e.g. Nuthall 2007) or to discern 'progress' over a single lesson, along with research evidence has led to the inspectorate recently adopting an ungraded observation approach.

The need to capture and quantify teachers' work is an inevitable consequence of the marketisation of education. Given the diversity and complexity of what teachers do, it thus makes sense for the market to narrow down the parameters of measuring teacher effectiveness to classroom practice, as this is something that policy makers perceive to be easier to quantify, hence the reliance on mechanisms like graded observation. Such practice is indicative of what Smith and O'Leary (2013, 246) have labelled as 'managerialist positivism', where the complexity of the teaching and learning process is superficially reduced to the presentation of quantitative performance data in order to satisfy the needs of a market that thrives on the production and comparison of such data. As is the case with the recent development of comparative international testing systems such as PISA for students, the search for 'best practice' has become a key driver for teacher assessment and development in England, with 'comparison not only possible but imperative' (Kamens 2013, 123). As Stevenson and Wood (2013, 44) have argued, it is the growing dependency on high stakes testing of the international education market that has been instrumental in 'transform[ing] teachers' labour into a product that can be quantified and measured'.

Although classroom observation is an activity used for multiple purposes, from research insights to performance management, it is its use as a mechanism of accountability through grading that has come to occupy centre-stage in England's colleges and schools in recent years. Whilst graded observations have been the subject of intense debate, one perspective which is rarely considered is that of the epistemological and methodological underpinning which informs how data from observations are understood and used by observers.

Performative lesson observations implicitly rely on a reductive worldview that tends to be underpinned by the notion that complex phenomena such as classrooms are best explained by separating out their constituent parts and analysing each separately. Observation frameworks based on the use of competency-based checklists are central to this form of observation as the process of pedagogy is often atomised into a series of individual variables, all of which are perceived to be identifiable and as such can be checked off against a predetermined list of features of a 'good' or 'outstanding' lesson. The 'observation scorecard' included in Appendix 1 (see below) provides a concrete example of such a metrics, competency-based model in practice from a large college in England.

Such competency-based models of observation, especially those that use a numerical marking/ranking scale, give the guise of adopting a 'scientific' approach to the measurement of classroom practice. As numbers have a 'scientific' quality to them, there can be a tendency for people to be less likely to question what they are deemed to represent. Criteria for an 'effective' lesson are presented in a prescriptive checklist and provided all these criteria are met then the lesson can be judged 'effective' as a whole. But this process is highly reductive insomuch as there is an assumption that if each individual criterion is present, then the whole must, by definition, be of the defined quality. It also assumes that all facets of the pedagogic process are clearly and quantifiably observable; as such the classroom is characterised as a simple, linear system where all causes and effects are understood and can be easily accounted for and coalesce to give a clearly defined whole. However, there is a rich tradition that characterises systems such as classrooms as being 'complex' i.e. non-linear systems where elements and interactions are rich and cannot be predicted in a simple, 'mechanistic' way.

Classrooms as complex phenomena: the quantifiable conundrum

Since the late 1990s an alternative view of teaching and learning has begun to attract significant interest, the idea of complexity theory (Morrison 2002; Davis & Sumara 2006; Mason 2008). Complexity theory maintains that many natural and social systems are not composed of simple, linear relationships but are instead formed from complex processes which cannot be understood by recourse to reductive analysis.

Cilliers (1998) describes complex systems as being characterised by large numbers of interacting elements, which are non-linear in nature meaning that large-scale causes can have small effects and vice versa. Interactions are also typified by negative and positive feedback loops. Negative feedback loops are characterised by processes which constantly lead the system back towards an equilibrium state, whilst positive loops lead to movement of the system to far-from equilibrium states. Complex systems are also 'open' meaning that they interact with the wider environment, making the borders of the system difficult to identify. Richardson et al (2007, 26) likewise describe complex adaptive systems in the following way:

> A complex (adaptive) system can be simply described as a system comprised of a large number of entities that display a high level of interactivity. The nature of this interactivity is mostly non-linear, containing manifest feedback loops.

Classrooms share many of the hallmarks of complex adaptive systems as they are nonlinear, which means that simple cause and effect understandings of interactions are problematic as they are both multiple and non-proportional. In addition, classroom interactions are not particularly predictable at any level of detail. For example, a simple reply to a student's question might have a fundamental impact on their conceptualisation of an area of subject knowledge, whilst a large-scale revision of a curriculum area may have a minor impact on student understanding. The permeable boundaries of a complex system (Cilliers 1998) are apparent in classrooms; the boundaries cannot be drawn by the walls as a whole series of external factors may impinge upon the ability and readiness of students to learn, and equally diverse factors impact upon the activity of the teacher, e.g. through curriculum decisions or government policy. One of the consequences of this difficulty in drawing boundaries to any particular system is that the system becomes *incompressible.*

Incompressibility is a crucial feature of a classroom system and makes any form of reductive observation process highly problematic. As Cilliers (2005, 13) observes:

> We have seen that there is no accurate (or rather, perfect) representation of the system which is simpler than the system itself. In building representations of open

> systems, we are forced to leave things out, and since the effects of these omissions are nonlinear, we cannot predict their magnitude.

Developing this argument, Richardson et al (2007) highlight that any model of an incompressible system would need to be as complex and as extensive as the system itself to be an accurate representation of it. Therefore, if the classroom is accepted to be a complex adaptive system, then any attempt to use a form of tick-sheet led observation would require that sheet to be as complex, and have as many factors, as are present within the whole of the system and its external links. To attempt anything less would begin to show a gross level of reductionism and hence oversimplification of the processes occurring within any given lesson. Richardson and Tait (2010) stress that any representation of an incompressible system will by definition be incomplete, though it may still be useful in helping us gain insights. Due to the incompressibility of complex systems, we are constantly involved in attempts to simplify what Biesta (2010) calls 'complexity reduction'; a crucial process for making experiences and systems intelligible. Therefore, we can still learn from observation and other forms of data capture in classrooms, but we have to accept that these insights remain incomplete and as such are not reliable indicators as performance orientated activities.

Incompressibility also leads to a discussion concerning 'local' versus 'non-local' knowledge. Non-local knowledge is that which has value over a broad range of different contexts and gives us the foundation for generalisable statements. Conversely, local knowledge is that which is highly contextualised and cannot be generalised to create universal statements concerning the system and its elements.

Therefore, a complexity view of the classroom leads to an argument that both reductionism and generalisability are highly problematic in relation to lesson observation. It strongly suggests that any use of observation for either internal performative purposes or by external bodies leads to a level of reductionism that can render its use meaningless, especially when relied upon as the sole, or main, source of evidence on which to base judgements about teachers' professional capabilities.

Research design

This paper draws on data from a year-long research project that took place from 2012-13 and was funded by the University and College Union (UCU) in England, the largest professional body representing members in both further and higher education. The focus of the project examined the use and impact of lesson observation on staff working in a range of contexts and institutions in the FE sector.

The research project adopted a mixed-methods approach involving quantitative and qualitative methods of inquiry. The rationale for a mixed-methods design was pragmatic and principled. It was pragmatic in the sense that developing as thorough an insight into lesson observation as possible was what drove the selection of research methods overall rather than any affiliation to a specific methodological paradigm. As Tashakkori and Teddlie (1998, 21) state:

> For most researchers committed to the thorough study of a research problem, method is secondary to the research question itself, and the underlying worldview hardly enters the picture, except in the most abstract sense.

Thus decisions about what data to collect, what were deemed to be the most appropriate and effective means of collecting the data, along with what to do with the data were 'dictated by the research question[s]' (Newman and Benz 1998, 15), the underpinning aims of the study and a commitment to the quality of the research.

The decision to use mixed methods was also principled in the sense that the study was conducted on the basis that neither a qualitative nor a quantitative approach can be considered superior to the other. For mixed methods researchers, 'the world is not exclusively quantitative or qualitative; it is not an either/or world but a mixed world' (Cohen, Manion and Morrison 2011, 22). Both methodological approaches have their strengths and weaknesses, as others have argued (e.g. Punch 2006) and 'even greater strength can come from their appropriate combination' (Gorard and Taylor 2004, 1).

An online survey, semi-structured interviews and focus groups were the main research tools used as part of a triangulated framework to address the project's research questions. Some of these questions were of an explicitly factual nature and thus leant themselves to a quantitative method of inquiry. Others sought to explore the lived experiences and perspectives of practitioners in the form of a narrative and so required a qualitative approach.

The sample for the first phase of the data collection (a web-based survey via SurveyMonkey) comprised UCU members, ranging from part-time tutors to senior managers. Approximately 4,000 respondents completed the survey (n = 3976) with over four fifths (86.4%) identifying themselves as lecturers/teaching staff, just under a tenth (9.1%) as middle/senior managers and the remaining respondents as 'other'. UCU FE membership was reported to be approximately 32,000 at the time the survey was circulated, thus there was an overall response rate of 12.5%. The second phase involved focus groups and individual interviews with thirty staff from a range of colleges across England, including UCU members and non-members. Purposive sampling was used to select the colleges to ensure a geographical spread, thus colleges were selected from the north, the midlands and the south of England.

The data analysis process began with the online survey, which contained quantitative and qualitative data. The quantitative data were analysed via SPSS and relied largely on the use of descriptive statistics, which were presented and discussed via frequency distribution tables and graphical displays (e.g. bar charts) in the study's report (UCU 2013). The qualitative data analysis combined steps outlined by Creswell (2003, 191-195) and Miles and Huberman's (1994, 9-12) 'three concurrent flows of activity' of analysis i.e. data reduction, data display and conclusion drawing/verification.

Findings and discussion

The impact of performative observations on teachers' professional practice

The survey's statistical data confirmed that observation had come to be regarded and indeed implemented largely as a mechanism for monitoring and controlling teacher performance with over four fifths (84%) of respondents indicating that their most recent

experience had been in the context of a quality assurance exercise as **Figure 1** below illustrates in participants' responses to that particular question.

Figure 1 – Contexts of lesson observation

As the responses in **Figure 1** demonstrate, the most common context selected by over two thirds (68.6%) of respondents was the Internal Quality Assurance (QA) scheme, where the lesson is evaluated and graded against the Ofsted 4-point scale. The context of 'external consultation' follows a similar approach and is typically used and referred to in common parlance as a 'Mocksted' by many organisations, where external consultants are employed to carry out observations across the institution with a view to simulating the experience of a real Ofsted inspection. When combined, the first three contexts listed in **Figure 1**, all of which adopt a similar performance management approach, amounted to over four fifths (84%) of responses. The data clearly emphasised a bureaucratic, reductionist view of observation where the complexity of teachers' work and its perceived impact on learning was being reduced to simplistic numeric data patterns within and across organisations.

The project's qualitative data contained repeated examples of practitioners drawing attention to how little input or control they had over the use of such performance management models of observation in their workplaces. The overriding impression was one of a top-down policy imposed by senior management, as the following comment from a survey respondent encapsulates:

> Observations in my college are little more than a means of managerial control. They call it "benchmarking" but it's just a made-up word for keeping teachers in their place. To me the purpose of the whole thing benefits management rather than us. We have no say over them whatsoever or input into the process at all. They don't make the slightest bit of difference in helping me to improve my teaching and only serve to undermine the morale of staff as a whole (Respondent 1456) [2].

[2] Interviewees are denoted by pseudonyms and survey respondents by identity numbers throughout this paper.

The 'control' referred to in this comment links to earlier discussion regarding the way in which the policy technologies of managerialism and marketisation have colonised teaching by demanding that teachers' work be accounted for quantifiably (Stevenson & Wood 2013). Performance management systems invariably rely on the production of statistical data that can be used for comparative analysis to measure the performance of individuals and institutions alike. In the case of lesson observations, the embodiment of this is the 'grade profile', which has become an established feature of accountability systems in FE in recent years and is relied on heavily by senior managers as a key audit tool with which to measure and compare levels of staff performance year on year internally and against national benchmarks. It is seen as a vital component of an institution's self-assessment for teaching and learning as others have argued (e.g. O'Leary 2013).

The grades from these observations are used to performance manage individual teachers, together with providing evidence for external inspection purposes, thus demonstrating Ofsted's hegemonic role in shaping senior managers' approaches to evaluating teaching and learning. The gathering and analysis of statistical data from annual graded observations is therefore seen as an essential part of internal audits for senior managers, which explains the reference to management as the main beneficiaries of observations in the quote above. However, this demonstrates a lack of engagement with the complexity of the processes behind the data which are generated and the possible loss of sight from extreme complexity reduction. As Wood (2014) has argued previously, the rise of 'dataveillance', the accumulation and tracking of data becomes more important than the individual to whom the data relates, and this was something of which the study's participants seemed particularly aware. In prioritising such activity, genuine opportunities for understanding and growing practice may well be lost in the attempt to generate data.

The perception of a significant majority of practitioners was that these observations failed to have any positive impact on their classroom practice or the improvement of the quality of teaching and learning. For example, the survey data revealed that three quarters of respondents (74.8%) disagreed that graded observations had helped them to improve as classroom practitioners. This level of disagreement (76.7%) was similarly reflected in responses to a question relating to whether graded observations had helped to raise the

standards of teaching and learning in their workplace. The majority of responses revealed an overwhelming discontent with the use of graded observations for teacher assessment and accountability purposes. These views were reinforced in the qualitative data as this small sample of comments from survey respondents and interviewees below illustrates:

> The regime of graded lesson observations is putting unbearable pressure on lecturers. It does not help develop good teaching and learning (34).

> Current graded system places undue stress on observees. They're seen in many colleges as a management exercise to satisfy external bodies (241).

> I don't see the value in a one-off, one hour graded observation that judges a teacher based on 0.12 % of the work they do (Isabel, senior manager).

These comments were indicative of the lack of value associated with graded observations by many participants, although there was one group whose views differed markedly from the majority. The only outliers amongst respondents were senior managers, who strongly defended the use of graded observations in their survey responses. Yet despite their support for this practice, there was a noticeable dearth of evidence to justify this position in their qualitative comments. Even in those instances where comments were broadly supportive of the use of graded observations, they tended to be accompanied by conditional statements emphasising the need to ensure that teacher development remained central to the process.

The qualitative data highlighted how performance management observations were invariably regarded as a perfunctory mechanism, with both observers and observees questioning their effectiveness and only senior managers defending their use. 'Box ticking' and 'jumping through hoops' were phrases that recurred frequently in the qualitative data when participants were asked to comment on the purpose(s) of observation in their workplace. This discourse revealed how the process was not valued by many teachers, but also how much of the time was spent measuring procedural aspects rather than actually measuring or improving the real quality of teaching and/or the overall quality of the

students' experience. Just under three quarters (73%) of survey respondents disagreed that graded observations were an important part of staff appraisal, with a similar percentage stating that they should no longer be used as a form of teacher assessment.

Such widespread disregard and dissatisfaction amongst practitioners was illustrative of the failings of performative approaches to observation rather than observation per se, along with the missed opportunities for professional learning that arose when an institution's approach was driven by a performance management agenda. Far from dismissing the value of observation outright, many of the study's participants acknowledged the instrumental role it had to play in fostering professional learning, especially peer-based approaches to observation which were repeatedly cited as being best suited to generating sustainable change and meaningful professional learning. It was therefore these approaches that should be at the forefront of an institution's use of observation and wider teacher development strategies; a point discussed in more detail towards the end of this paper.

A large proportion of the project's qualitative data pointed to how the use of graded observations had given rise to a network of interconnected, counterproductive consequences for teachers, highlighting the predominant perception among many participants that observation was deemed problematic rather than productive. The research data repeatedly revealed how performative models of observation were perceived as having constraining and negative effects on teachers' professional identities. Examples of this included the labelling of teachers according to their grade, despite the claim by senior managers that it was the 'learning' that was being assessed rather than the teaching, along with increased levels of anxiety and stress.

Evidence of implicit and explicit labelling of teachers occurred across data sets and was commented on by a wide range of participants from hourly paid tutors to senior managers. Many of these comments drew attention to some of the deleterious effects of observation policies individually and collectively as Elizabeth, a basic skills tutor, implies here:

> The grading of observations is divisive – we are given tables of how many people got which grade – it has almost become unhealthy competition – it's unnatural too.

> Personally I hate the process though I get good grades. I live in fear of failing next time (Elizabeth).

Elizabeth's account was by no means an isolated one. As her comments reveal, the labelling effect of grading in her workplace had threatened to undermine collaborative learning amongst colleagues, instead giving rise to 'unhealthy competition' amongst colleagues and adversely affecting levels of morale and cooperation as she later went on to describe. Her account evokes Ball's (2003) seminal critique of performativity and how it is inherently divisive as it engenders a culture of individualism and professional identities that fail to appreciate the importance of collegial loyalty and cooperative support. Once again, this exemplified a failing of these performative approaches to observation. By ranking teachers and pitting their classroom performances against one another individually, not only was this resulting in the pathologisation of teachers' practice, but valuable opportunities for collaboration and reciprocal learning amongst teachers were being missed. Situated in the context of 'high stakes' assessment, Elizabeth's analysis also echoes Gipps' (1994) work, who found that with high school students normative grading threatened collaborative learning by causing unhealthy competition and impacted negatively on levels of motivation.

'Stress' appeared repeatedly in the qualitative data in reference to graded observations. Taking the textual comments from the survey as one example, over a quarter of the 1619 responses included the word 'stress', often in conjunction with other terms such as 'anxiety' and/or 'pressure'. Many participants associated the whole experience of lesson observation, particularly for performance management purposes, with a set of predominantly negative emotions. This was something that was not restricted to the act of being observed but occurred in the lead-up and post-observation period, and in some cases had more far-reaching consequences for individual teachers' health and well-being.

Performative models of observation were perceived as undermining professional trust. There was the suggestion that the increase in the frequency with which mechanisms of accountability and surveillance such as graded and unannounced observations were being used 'leads to an ethos where FE staff feel that their professionalism is not respected' and that they could not be trusted to do their jobs without being subjected to 'constant scrutiny'

as one focus group participant commented. There were calls for a greater level of self-regulation, as the two survey comments below demonstrate:

> Although other jobs have appraisal systems to monitor performance (as do we as teachers) they don't have the constant scrutiny that teachers have. If the government consider us as "professionals" how come they don't believe we can self-regulate and yet other areas outside teaching can? (416)

> I hate being observed. 30 years in teaching and still can't be trusted. The attitude seems to be 'You are only as good as your last observation'. What other profession requires continual monitoring on this scale? Name one! (501)

Capturing the complexity of classrooms and teachers' work

Like any other form of data collection, observation has its strengths and limitations. Participants repeatedly drew attention to the perceived inadequacies of observation as a means of attempting to capture the complexity of classrooms and teachers' work, particularly when relied upon as the sole or main data source. In their qualitative comments, some teachers highlighted the difficulty of observing classroom environments given the number of factors and issues which simultaneously occur at any given point in time. This was illuminated in their reflections on the complexity of the classroom and the inadequacy of observation as a tool to capture it:

> We work in a hyper-complex environment in terms of teaching and learning. I mean there are so many aspects to a teacher's job that you can't possibly get an insight into just by observing. Surely it's got to be a more holistic thing, hasn't it? (Penny, Director of Quality).

Others manifested an understanding of the permeability of classroom boundaries and the fact that learning in any one lesson was part of a trajectory over time. As argued earlier, the boundaries of the classroom extend beyond its spatial and temporal elements (Cilliers 1998). To see learning as a process that occurred within a single lesson only served to demonstrate a poverty of the complexities involved. Indeed, there appeared to be a

considerable mismatch between a complexity view of pedagogy as understood by some teachers and a reductive notion of learning, which underpinned the perceptions of those observers only interested in conceptualising and compartmentalising learning as it appeared at a given moment:

> There are many variables in the numerous complex relationships between teacher and students. I've been graded low when an observer selected a couple of learners to interrogate who were struggling with a concept. The learning for those students was achieved over a number of weeks and it was not a problem for me that they hadn't grasped it in that particular lesson. However it was perceived as a problem by the observer (Richard, engineering lecturer).

Here, teachers are showing a natural understanding of the classroom as an open, complex system, which is constantly altered and impinged upon by outside influences. Contrary to this view, the episodic nature of annual graded observations must be considered a contributory factor in a compartmentalised conceptualisation of learning, though as Maryam, an experienced observer, remarks below, individual observers still have an agentic role to play in the process. However, as she suggests, some seem less willing to conceptualise classrooms as complex adaptive systems than others:

> I don’t think graded or unannounced observations give an accurate picture of what happens day-to-day in the classroom. There are too many variables to consider and, however much training is given to observers, it seems that some observers have their ‘own agenda' when observing … The fact that every group is different in FE and that what may be considered excellent practice with, say, a Level 3 Access to Higher Education class may not work at all with an auto-engineering group of challenging teenagers does not seem to be considered by all observers unfortunately as some seem to have a very fixed view of learning and what they’re looking for rather than going in with an open mind and focusing on what they see. I’m always open to talking to staff about the bigger picture but I know several of my fellow observers aren’t.

There was also evidence that some participants saw the reduction of their teaching through observation to a single grade or number as being wholly unreflective of the nature of the task with which they were engaged. Consequently, it seemed that some teachers identified the process of observation not only as being unhelpful, but also as one not to engage with as it no longer resonated with their perceptions of the incompressibility of practice:

> Of a 37 hour a week, your average full-time member of staff will teach 24 hours and plan and mark the rest of those hours. There's a little bit of time where they're doing some course management but when it comes to appraising their performance we ask them 'what grade did you get' in your annual observation and that's about it (Ian, Head of Department).

Ian's remarks also epitomised the way in which observation has become fetishised in FE in recent years insomuch as it has taken on the status of a reductive, one-size fits all mechanism with which to assess teachers' practice performatively, but also to diagnose their CPD needs. This reductionism clearly resonated with many practitioners, particularly the more experienced ones. As Fiona, a business studies lecturer with over 30 years' experience commented, 'I resent and absolutely hate being reduced to a number after being closely watched by someone hiding behind a clipboard for an hour.'

One of the clearest themes to emerge across multiple data sets and participant groups was how the current reliance on annual graded observations as a means of measuring practitioners' professional capabilities was considered not only reductive but also inequitable practice. Although the data revealed a broad consensus among participants that accountability was an inescapable element of what it meant to be a teaching professional in FE, the reliance on episodic observations as the main evidence on which to base summative judgements about their professionality provoked a high level of opposition. Not only did such practice fail to capture the breadth and complexity of participants' work, but the shortcomings about its validity and reliability as a method of assessment in itself raised serious question marks about its fitness for purpose.

Participants at all levels of seniority acknowledged the need to move towards a more sophisticated, complex model of trying to capture what it is teachers do and the impact of their work; one in which other relevant sources of data could be drawn on to inform and supplement evidence of teacher performance gathered during observations. Given the partial insight into practice provided by episodic observations, participants suggested extending it beyond the lens of lesson observation and incorporating other sources of data into the assessment process so as to provide a more valid, reliable and triangulated evidence base for assessment. Examples of these suggestions are included in **Figure 2** below, which advocates a multi-dimensional model of teacher evaluation.

Figure 2 – Multi-dimensional model of teacher evaluation

On a practical level, how such different data sets might be combined and implemented into a coherent and workable framework of assessment remains unclear at present. This undoubtedly presents a significant challenge for any organisation and would benefit from further research. However, what does appear clear from these findings is that observation, when used in isolation, is an inadequate method of analysing the complex interactions of classrooms and teachers' work. Nevertheless, it can still be reconceptualised and relocated as a useful process when seen as part of a greater whole. In the words of Richardson and Tait (2010, 92-93):

> Just because a complex system is incompressible it does not follow that there are (incomplete) representations of the system that cannot be useful – otherwise how could we have knowledge of anything, however limited? Incompressibility is not an excuse for not bothering.

Richardson et al (2007) suggest that if complexity thinking is used as an epistemology (i.e. as the basis for knowledge claims, in this case understanding classrooms), then any analysis of complex adaptive systems requires consideration from a number of perspectives. Each perspective will be incomplete, and the system will never be understood in its entirety. However, by using a range of data collection methods, different elements of the system can be captured and will allow for the generation of multiple-perspective understandings. In

doing this, lesson observation becomes merely one of these perspectives, requiring us to look for other forms of evidence, such as the work produced by students, the videoing of lessons allowing us to deconstruct and understand the system in different ways, and the use of interview techniques to uncover, albeit imperfectly, the cognitive elements of student learning. Obviously, using a number of approaches to understand learning in a lesson takes time and this may not be feasible on a regular basis. Nonetheless, we need to begin to think inventively concerning the spectrum of information we can collect from a lesson that allows us to begin to triangulate and gain multiple perspectives on learning that has taken place. Lesson observation is thus an important element of such a mix but not enough on its own.

In summary, though the comments above came from participants across a range of institutions and in differing roles, what knitted their personal narratives together was their shared experiences of how performance-driven models of observation had oversimplified the complexities of classrooms and the teacher-student relationship under the guise of making teaching more transparent and measurable. What emerged from these narratives was a picture of how observation had thus become colonised as a simplistic, superficial tool that attempted to explain complex processes by means of converting them into 'simple figures or categories of judgement' (Ball 2003, 217). As Richard's comments above imply, context, individualised learning and the iterative nature of the dynamic between teachers and students were eschewed as confounding variables in the application of observation as a standardised instrument of measurement.

Beyond 'observation' and towards practitioner investigation: concluding thoughts

The above discussion suggests the need for a fundamental reconsideration of the place of observation in classroom settings. Its established use as a performative tool seems to have reached the threshold of its sustainability, as it has come to be relied upon inappropriately as a reductive tool to analyse complex processes in a non-linear environment. As some of the findings demonstrate, this has given rise to a plethora of counterproductive consequences that threaten to impede future advances in teacher evaluation and learning.

Not only has the impact of performative observation reached its threshold, but it has become an obstacle to aiding continued teacher learning and improvement. It is paradoxical

that part of the original justification for introducing graded observations was to identify and remove poor teachers from the profession, yet they have proven to be ineffective means in doing so as the performance element has led to a level of inauthenticity that has compromised the very core issues of validity and reliability of assessment, not to mention professional trust. If managers have to rely on one-off, snapshot observations of staff to be able to assess their professional capabilities, then arguably they are not managing their staff effectively in the first place. That is something that can only be achieved through sustained relationships, with managers observing and talking to their staff on a regular basis, along with drawing on a range of complementary evidence. Besides, such evidence should be drawn on over a period of time rather than as a 'snapshot' in order to create deeper understandings. This can be framed as an argument concerning the degree of complexity reduction (Biesta 2010) involved in discussing teaching and learning. Any process of analysing and considering change will involve such reductions as argued by Richardson and Tait (2010). However, the degree of reduction involved is crucial. If we accept, as they do, that only partial views are possible, then observation becomes a useful diagnostic and formative tool in helping the emergence of new practice. However, acute reduction, as in the collapsing of practice to summative, numeric identifiers and 'best practice' narratives may be deemed to distort the complexity beyond any useful point.

So, what are the ramifications for the future use of lesson observation in an educational context? We would argue that it would be positively utilised in a wider reformed view of teacher growth, where teachers see continuous classroom investigation and data analysis as part of their core role as professionals. There is a growing bank of evidence in the areas of coaching, mentoring and lesson study, for example, that serves to illustrate how lesson observation can be reconceptualised and repositioned as an opportunity for practitioner investigation rather than the narrow lens through which it is currently conceptualised in education as an instrument of teacher assessment (e.g. Cajkler & Wood 2016; Lofthouse & Hall 2014; Lofthouse & Wright 2012).

The findings from the study discussed in this paper make a strong case for arguing that the application of observation in an educational context needs to be underpinned by a supportive approach to teacher learning and growth rather than the current reductive

models that invariably operate on punitive principles. Professional trust must be at the core of any such approach rather than suspicion and distrust. As Darling-Hammond (2014) argues, teacher evaluation needs to be part of a teaching and learning system that supports continuous improvement for teachers individually and as a collective community of practice. Fostering collaborative learning amongst teachers is much more likely to support student achievement than divisive, ranking exercises.

No longer should observation be regarded and implemented as a predominantly summative assessment tool or disciplinary mechanism, but instead as a method of inquiry that contributes to a continuous professional dialogue based on self-reflection, action research, feedback, peer coaching and experiential learning. Only with such developments can the profession begin to reclaim observation as an empowering tool in the future growth of teachers. However, this will only happen if teachers are encouraged to experiment and to expose their practice to the eyes of others without the fear of punitive surveillance systems, but instead joined by a common pursuit of furthering their understanding of the complex processes of teaching and learning. And in turn this is only likely to occur once there are significant shifts in education policy more broadly, the way in which teachers' work is regarded and their ability to exercise greater agency over what they do.

Acknowledgement
The UCU (2013) report referred to in this paper was a national project funded by the University and College Union. The authors wish to thank UCU for agreeing to the sharing of some of the project data.

References

Ball, S. J. 2003. The teacher's soul and the terrors of performativity. *Journal of Education Policy*, 18(2), 215-228.

Ball, S. J. 2012. *Global Education Inc: New Policy Networks and the Neo-liberal Imaginary.* London: Routledge.

Biesta, G. 2010. 'Five theses on complexity reduction and its politics.' In D. Osberg and G. Biesta (eds.) *Complexity Theory and the Politics of Education*. Rotterdam: Sense Publishers, 5-14.

Cajkler, W. & Wood, P. 2016. Adapting 'lesson study' to investigate classroom pedagogy in initial teacher education: what student-teachers think, *Cambridge Journal of Education*, 46(1), 1-18.

Cilliers, P. 1998. *Complexity and Postmodernism: understanding complex systems.* London: Routledge.

Cilliers, P. 2005. 'Knowing complex systems.' In K.A. Richardson (ed.) *Managing Organisational Complexity: Philosophy, Theory, and Application*. Charlotte: Informational Age Publishing, 7-19.

Cohen, L., Manion, L. & Morrison, K. 2011. *Research Methods in Education* – 7th Edition. London: Routledge.

Creswell, J. 2003. *Research design: qualitative, quantitative, and mixed approaches* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Darling-Hammond, L. 2005. Teaching as a profession: Lessons in teacher preparation and professional development. *Phi delta kappan*, *87*(3), 237.

Darling-Hammond, L. 2014. 'One Piece of the Whole: Teacher Evaluation as Part of a Comprehensive System for Teaching and Learning,' *American Educator* 38, no. 1 (Spring 2014): 4–13.

Davis, B. & Sumara, D. 2006. *Complexity and Education: Inquiries Into Learning, Teaching, and Research.* Mahwah, N.J.: Routledge.

Department for Education (DfE) 2010. *The importance of teaching* – Schools' White Paper. London: DfE.

Exley, S. 2014. 'Why the score is not even on graded lessons', *TES Magazine,* https://www.tes.co.uk/article.aspx?storycode=6443156 Accessed 11/01/2016.

Gipps, C. (1994) *Beyond Testing: Towards a Theory of Educational Assessment*. London: Falmer Press.

Gorard, S. & Taylor, C. 2004. *Combining Methods in Educational and Social Research*. Berkshire: Open University Press.

Huddlestone, P. & Unwin, L. (2013) *Teaching and Learning in Further Education – Diversity and change,* 4th Edition. London: Routledge.

Kamens, D. 2013. 'Globalization and the Emergence of an Audit Culture: PISA and the search for 'best practices' and magic bullets', in Meyer, H-D and Benavot, A. (Eds) *PISA, Power and Policy: the emergence of global educational governance*. Oxford, Symposium Books.

Lofthouse, R. & Hall, E. 2014. 'Developing practices in teachers' professional dialogue in England: using Coaching Dimensions as an epistemic tool'. *Professional Development in Education*, 40(5), 758-778.

Lofthouse, R. & Wright, D. 2012. 'Teacher education lesson observation as boundary crossing'. *International Journal of Mentoring and Coaching in Education*, **1**(3), 89-103.

Mason, M. 2008. *Complexity Theory and the Philosophy of Education.* Chichester: Wiley and Sons Ltd.

Meyer, H-D & Benavot, A. (Eds) *PISA, Power and Policy: the emergence of global educational governance*. Oxford, Symposium Books.

Miles, M. B. & Huberman, A.M. 1994. *Qualitative Data Analysis*, 2nd Edition. Newbury Park, CA: Sage.

Morin, E. 2008. *On Complexity.* New York: Hampton Press

Morrison, K. 2002. *School Leadership and Complexity Theory.* London: Routledge.

Morrison, N. 2015. 'Ofsted to scrap graded lesson observations in FE', *TES Online*. Available at: https://www.tes.com/news/further-education/breaking-news/ofsted-scrap-graded-lesson-observations-fe. Accessed 14/01/2016.

Newman, I. & Benz, C. 1998. *Qualitative-Quantitative Research Methodology – Exploring the Interactive Continuum*. Carbondale, IL: Southern Illinois University Press.

Nuthall, G. 2007. *The hidden lives of learners*. NZCER Press.

O'Leary, M. 2013. Surveillance, performativity and normalised practice: the use and impact of graded lesson observations in Further Education Colleges. *Journal of Further and Higher Education,* 37(5), 694-714.

O'Leary, M. 2014. 'Power, policy and performance: learning lessons about lesson observation from England's Further Education colleges'. *Forum*, 56(2), 209-222.

O'Leary, M. 2015. 'Breaking free from the regulation of the State: the pursuit to reclaim lesson observation as a tool for professional learning in Further Education', chapter in Daley, M., Orr, K., & Petrie, J. (eds) *Further Education and the Twelve Dancing Princesses,* London: IoE Press.

Power, M. 1994. *The Audit Explosion*. London: Demos.

Punch, K. 2006. *Developing Effective Research Proposals* – Second Edition. London: SAGE Publications.

Richardson, K.A.; Cilliers, P. & Lissack, M. 2007. 'Complexity Science: A 'Gray' Science for the 'Stuff in Between' in *Thinking Complexity: Complexity and Philosophy volume 1*, Cilliers, P. (Ed.). Mansfield, USA: ISCE Publishing, 25-35. Retrieved from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.94.7038&rep=rep1&type=pdf) 21/10/2014.

Richardson, K.A. & Tait, A. 2010. 'The Death of the Expert?' *E:CO*, 12(2), 87-97.

Smith, R. & O'Leary, M. 2013. New Public Management in an Age of Austerity: Knowledge and Experience in Further Education, *Journal of Educational Administration and History*, 45(3), 244-266.

Stevenson, H. & Wood, P. 2013. Markets, managerialism and teachers' work: the invisible hand of high stakes testing in England. *The International Education Journal: Comparative Perspectives,* 12(1), 42–61.

Stronge, J. H., Ward, T. J., & Grant, L. W. 2011. What makes good teachers good? A cross-case analysis of the connection between teacher effectiveness and student achievement. *Journal of Teacher Education*, *62*(4), 339-355.

Tashakkori, A. & Teddlie, C. 1998. *Mixed Methodology – Combining Qualitative and Quantitative Approaches*. London: Sage.

University and College Union (UCU) 2013. *Developing a National Framework for the Effective Use of Lesson Observation in Further Education*. Project report, November 2013. Available at: http://www.ucu.org.uk/7105.

Wood, P. 2014. 'Teacher Professionalism: subverting the society of control.' *Forum,* 56(2), 223-234.

Appendix 1 – Observation Scorecard from Darkside College

SCORECARD

Planning 8%	**grade**	Indiv grade PIs	**Learning 44%**	**grade**	Indiv grade PIs
• Outcomes are relevant, at the correct level/aligned to the syllabus/set in appropriate context and documentation • Know your learners. ❖ learner needs against starting point ❖ effective use of ALS • Application of learner profile. ❖ to include accounting for learners prior knowledge/skills • Productive use of time. ❖ chunking learning			• Attendance and punctuality • Learners are sufficiently stretched and challenged to narrow gaps in achievement • Variety of activities to promote full engagement • Learners show evidence of development of English/Maths and F/S • Independent and committed learners • Learners demonstrate progress in their learning session • Skills and language to learn • Peer support, collaboration and respect. • Impact/relevance in the application of equality, diversity and differentiation. ❖ Citizenship and diversity ❖ Diversity and inclusive practice ❖ Discovering and celebrating diversity ❖ Diversity and employment • Adding value ❖ raising aspirations ❖ employability skills ❖ distance travelled, ❖ life skills ❖ experience of work ❖ sustainability. • Independent learning outside the learning session		
Teaching/Underpinning knowledge 28%	**grade**	Indiv grade PIs	**Assessment 20%**	**grade**	Indiv grade PIs
• Innovative and reflective approach to teaching and learning • Embedding equality and diversity through curriculum delivery, taking advantage of naturally occurring themes • Tutor / Trainer enthusiasm, support and guidance • Subject expertise, including Maths, English and F/S • Opportunities created to develop Maths, English and F/S • Health and safety/safeguarding • Innovative use of resources			• Effective use of questioning, stretch & challenge taking account of individual needs * how, why, what • Variety of assessment strategies to meet individual needs formative/summative • Learner progress against outcomes • Feedback informs learners how to improve • Learner reflection ❖ Learners understand what they have to do to improve their skills and knowledge		
Outstanding 90% and above, Good 80%-89%, Requires improvement, 65%-79%, Inadequate 64% and below			**Overall grade**		