Observation Manuals as Lenses to Classroom Teaching: Pitfalls and Possibilities

Kirsti Klette & Marte Blikstad-Balas

Printed in: *European Educational Research Journal*, 1474904117703228, vol 17, issue 1, 2018

URL: http://journals.sagepub.com/doi/abs/10.1177/1474904117703228

Abstract

The aim of this paper is to discuss the role of coding and observation manuals in classroom studies. While observation manuals have been a part of the methodological toolkit for measuring various aspects of instruction for decades, the field has also been suffering from "paradigm wars", fragmentation and local production of instruments. Common frameworks for investigating teaching are needed, including observation instruments for teaching that are both generic and subject specific. Such common tools for research developed within an integrated methodological design could help researchers make progress in aggregating knowledge about the impact of different teaching approaches across settings and subjects. This article serves as one such integrative mechanism by summarizing and reviewing existing manuals targeted towards developing knowledge for and in teaching. The analysis provides status, overview and focus of the different observation manuals; additionally, the article discusses how recent developments in instruments and coding procedures might provide increased rigour and a shared vocabulary to talk about teaching. We discuss both pitfalls and possibilities of coding manuals, and argue that if used in a reflexive manner, coding manuals can provide a common language and vocabulary when talking about – and researching – classroom teaching and learning.

Keywords: Coding manuals, video studies, comparative classroom research

Introduction

Observation plays a key role in providing information about instructional practices, both as a research method and as part of teachers' continuous learning and development. Teachers might observe each other, principals might observe their teachers, and teacher educators and teacher mentors might observe student teachers. Observation also has a longstanding tradition as a central part of the educational researchers' methodological repertoire for understanding instructional practices within and across different subjects and learning sites, and it thus provides a toolkit for comparative classrooms studies (Alexander, 2001; Clarke et al., 2006) and comparative didactics (Hudson and Meinert, 2011). Even though observation holds a strong position for understanding teaching, the vast majority of observations are based on unstandardized, informal and un-validated instruments (Stuhlman et al., 2010). They can be based on empirical evidence and experimental studies, they can be based on theoretical assumptions or they may simply be a reflection of personal preferences (Stuhlman et al., 2010). The increased use of video recordings as data in educational sciences – and the methodological affordances of video – calls for a renewal of traditional debates on the strengths and limitations of standardized observation protocols. Seminal large-scale studies such as the *TIMSS video study* (Stigler & Hiebert, 1999, Kuger & Klieme, 2016) and the *Measures of Effective Teaching project* (BMGF, 2012), together with recent developments within video technology (e.g. miniaturized cameras, remotely controlled recordings and software tools for analysing large data sets) has generated significant research and development work on observations systems, including systematic testing of the different observation manuals-.

Video observation is an increasingly popular method by which to analyse teaching and learning due to benefits such as the ability to capture both the teacher's and students' perspective "in one package" (Fischer and Neumann, 2012), to decompose teaching practices into smaller entities (Klette, 2009) and the possibility of approaching the same segment of recorded teaching with different analytical foci (Blikstad-Balas & Sørvik, 2015; Jewitt, 2012). Rich and Hannafin (2009) emphasize how the combination of video data and newer video annotation tools makes it possible to record, review, analyse and synthesize different instructional practices. This can increase the overall credibility of the observations, not only because the initial observer may see the same segments multiple times but also because traditional quality assurance strategies such as reliability testing , member checking and secondary analyses are enabled in new ways. Video data support transparency and explicitness, making codes and coding systems the object of joint discussions and

verification, including the development and validity and reliability testing of the different codes. Thus, video recordings have contributed to a renewed interest in observation designs and observation manuals, including the systematic testing of the different observation measures.

The purpose of this article is to compare and contrast available coding manuals as possible lenses and a basis for developing a shared vocabulary in the research of classroom teaching and learning. A key question we address is how codification and observation manuals, developed within the shared experience of video studies, contribute to increased transparency and the advancements of a shared common language targeted towards understanding teaching and learning.

We begin in Section I by discussing the role of codification and how codification and classification, explicitly and implicitly, are part of all educational empirical research. We argue how common coding manuals might contribute to (i) a strongly needed common vocabulary when studying teaching and learning, (ii) decompose teaching and learning into "studyable" segments, and thus (iii) facilitate comparative analyses across contexts and classrooms. The latter point is especially pressing as specific elements of classroom teaching and learning are considered generic features, what Tyack and Cuban (1995) and others describe as "the grammar of schooling", while simultaneously being contextual and culturally specific, embedded in local and "cultural scripts" (Stigler and Hiebert, 1999). Analyses, we argue, must therefore be explicit and transparent, open to adaptions, additions and comments, thus making it possible to compare classroom instruction across settings, subjects and contexts. In Section II we review some of the existing manuals, discussing similarities and differences across them, including theoretical grounding and views of learning. In the last and third section, we summarize key challenges of using common coding manuals and argue that coding manuals could serve as one of several mechanisms by which to establish a common language and vocabulary when talking about classroom teaching and learning

I The role of codification in classroom studies

The process of sorting and categorizing data in one way or another (e.g. classifying, labelling, coding) is a part of all empirical research, regardless of whether the researchers actively verbalize and identify their categorization into a set of codes (Hammersley, 2007; Maxwell and Chmiel, 2014). When employing observation instruments, the codes are usually clearly defined, either before the process of gathering data begins (for instance, if one collects data in a classroom using an observation scheme by ticking off different activities on a pre-defined list), or they can be developed and refined

during a first cycle of analysis, after identifying emerging trends in the data and deciding what types of codes appear to be worthwhile. According to Saldaña (2013: 3), "A code in qualitative inquiry is most often a word or short phrase that symbolically assigns a summative, salient, essence-capturing, and/or evocative attribute for a portion of language-based or visual data." Thus, in simple terms, coding is about identifying and labelling segments of data. The term *data* can encompass everything from transcribed interviews, to field notes, videos, documents or photographs, but in this article we are primarily concerned with observations or video data of naturally occurring teaching and learning in classrooms (as opposed to, for instance, interventions).

In methodological sections of classroom studies, and even in introductory publications about how to conduct research, the process of coding is sometimes described as a straightforward sorting of data into different categories. However, coding is more than a way of sorting data; it is a transformation of data from one form into another, wherein certain perspectives are bound to be systematically foregrounded and emphasized while others are not. Thus, codification changes not only how the data is organized into smaller entities but also how it is perceived as a whole. Foray has described codification as follows:

> Codification – the translation of holistic aural or pictorial expressions into symbolic content stripping away their individually expressive character – has broader cognitive implications than the simple improvement of transformation transfer and storage processes. In particular, codification shifts language from aural to the visual domain, making it possible to arrange and examine knowledge in different ways. (Foray, 2001: 1558).

Foray (2001; 2010) and others (Foray and Hargreaves, 2003; Nelson, 2000) define knowledge codification as a condition for production and accommodation of knowledge because it optimizes cognitive opportunities and knowledge transfer. Analysing knowledge growth and knowledge production across rather different disciplines such as logistics, health, education and leadership, Foray (2001) argues how poor codification of the education sector makes it difficult to produce "learning programs" or codified instructions that can be made the subject of comment and addition by practitioners and researchers. Nelson (2000) takes this argument even further when he claims that the understandings we do have about the principles of good teaching are the same principles we have known for generations. As he emphasizes, it is "…not clear that we know more than 100 years

ago", and this situation "stands in sharp contrast with other arenas of human know-how, like information processing, and communication or transport" (Nelson, 2000: 122).

The lack of shared practices and a common knowledge base is problematic because it leads to teachers systematically working primarily on their own — and the process of generating new professional knowledge thus becomes a personal and subjective one. D. Hargreaves (2000) describes this well when he claims that teachers are artisans who primarily work on their own, which also means that their repertoire of teaching is developed through trial-and-error. Thus, as Hargreaves (2000) emphasizes, teachers spontaneously go along tinkering in their classrooms – thereby making tinkering their main mechanism for knowledge generation. In a sense, codification can be seen as the opposite of personal tinkering, as it is a condition for the production and accommodation of knowledge. Codification enables researchers to reduce the richness of, for instance, classroom instruction and talk about allocated observed behaviour that has been identified following a set of well-defined and agreed-upon codes. Obtaining the coding system to ensure that the meaning of the analysis is the same between coders enhances the validity and certainty of findings, Morse (2015) argues. Although this certainly involves reduction (Blikstad-Balas, 2016; Snell, 2011), reduction is also what enables systematic comparison and accommodation of knowledge. In line with Grossman and MacDonald (2008), we therefore argue that if the field of teaching and learning is to move forward, there is a need to develop programmatic research that addresses a set of questions over time, including a range of common tools (e.g. codes/coding manuals) and approaches for making progress in answering those questions.

Much qualitative research on teaching and learning has been conducted with local stipulation of what certain concepts include and specific operationalizations of how these terms can be observed (e.g. what we have described as 'the local production of instruments'). As argued by Blikstad-Balas (2014), this can be problematic for educational scientists because we might wrongly assume that different studies are addressing the same phenomenon (and vice versa). For instance, studies of "classroom discourse" might emphasise rather different aspects of the discourse such as frequencies of utterances; participation patterns, turn taking and content involved in their analysis while at the same time using the same label — classroom discourse. Thus, if the differences between what the studies actually address are not explicit, one might wrongfully assume they all measure or attempt to measure the same phenomenon.

The richness and complexity of most topics studied in educational research makes it compelling to develop new sets of codes that suit a particular study. Indeed, in many studies it may be necessary to do so. However, we argue, in line with Foray (2001), that the educational sector in general suffers from poor codification in the sense that educational researchers often lack a common set of codes to refer to when investigating key instructional practices. In the following, we will elaborate what we identify as the three most important benefits of employing common observation instruments in studies of teaching and learning practices in the classroom.

A common technical vocabulary for describing instruction

A key benefit of using common observation instruments in studies of teaching is that such instruments may provide a common technical vocabulary for describing instruction. Instruction is complex, and while they may be reductionist, codes of instruction can be a way of ensuring that certain aspects of teaching are verbalized, made explicit and referred to in a more consistent way. Over and over again, scholars (e.g. Cohen and Moffitt 2009; Dreeben, 2005; Grossman and McDonald, 2008; Lortie, 1975) have pointed to the lack of a common vocabulary as a key hindrance in the development of knowledge for and in teaching. A common vocabulary, and a common set of codes, could serve as a crucial resource to teachers and researchers scholars: "a common language with which to identify, investigate, discuss, and solve problems of teaching and learning—and thus the elements of common professional knowledge and skill" (Cohen and Moffitt, 2009, p. 5). Without a common vocabulary grounded in a shared knowledge base, they argue, it is difficult to build knowledge relevant for practice and research. By "common vocabulary", we do not mean that a given manual can be applied as a generic "swiss knife" across different communities without careful adaption and adjustments including rigorous training and discussion of what the common technical language actually refers to. Instruction is a complex object of study, and any manual attempting to measure it in precise or in nuanced ways is bound not to be self-explanatory. This is also why so many observation manuals actually require training and certification. While complete consensus on what specific words refer to is utopic, stipulating and defining features of instruction, as many observation manuals attempt to do through a common technical vocabulary and supporting rubrics, can enhance validity and reliability (Morse, 2015) and improve credibility of research (Blikstad-Balas, 2014). While no set of codes could ever capture everything that matters in the classroom, a common technical language for instructional practices can be a great starting point for talking about teaching.

For instance, the Protocol for Language Arts Teaching Observations (PLATO), a classroom observation protocol designed to capture the most important features of English/Language Arts (ELA) instruction (Grossman et al., 2013), provides good examples of the power of a common technical vocabulary. Originally, this manual was developed to differentiate between more and less effective teachers in English Language Arts classrooms. Currently, PLATO is also used to capture instructional practices in mathematics and as a professional development tool in support of teachers' use of rigorous, research-based teaching practices. This is possible precisely because of the technical vocabulary that PLATO makes available. Observation instruments made to assess instruction rely on describing different elements of instruction as clearly as possible. These definitions will influence both "what counts" as evidence of different instructional practices and how these practices can be made the object of discussions and analysis amongst professionals (see for example Sun and van Es 2015) and researchers (BMGF, 2012; Grossman et al., 2013).

A clear benefit of manuals such as PLATO, Classroom Assessment Scoring System (CLASS) (Pianta et al., 2008) and the Framework for Teaching (FFT) (Danielson, 2013) is that they go beyond the question of *what* happens in observed lessons to address *how* different instructional activities are performed (see also Groninger et al., 2012; Klette, 2010). These manuals differentiate between high-end and low-end evidence of instructional practices – and make the nuances between these categories explicit. Instead of only looking at whether or not a teacher uses, for example, modelling in his or her teaching, observation manuals such as PLATO differentiate between degrees of modelling. This kind of detailed differentiation illustrates a shift from discussing modelling *per se* to discussing the differences between no modelling, partially successful modelling and higher level evidence of successful modelling. In the case of PLATO, this is done with a four-point scale that says something about the amount of evidence observed concerning a specific instructional practice. The vocabulary of PLATO makes it possible for researchers to discuss aspects of, for example, modelling within an explicit frame of reference. With video recordings, it is also possible to see the same segment with peers and discuss the links between the actual empirical data and the labels added to it via the coding manual. By doing so, the vocabulary provided by the manual will also become a part of the common discourse between peers working within the same area.

Applying a defined set of codes verbalized with a common technical vocabulary links the analyses of instruction to specific and explicit criteria, thus making the process of analysis transparent and

explicit. The transparency is strengthened through detailed descriptions of how different features of instruction are assigned different codes. This makes it possible for the scientific community to assess both whether the codes make sense and whether the inferences drawn from the data appear plausible. It may also increase consistency between studies. As emphasized by Hammersley (2010), researchers should strive to present and represent data to their audiences. Presenting a well-developed framework that explicitly describes how classroom data is categorized in a given study can make this crucial step in credibly communicating research more manageable. Critical re-examination is also facilitated with a common set of codes and a joint vocabulary of instruction, and often such re-examination is a natural part of validating codes and assessing inter-coder reliability. Thus, a common technical vocabulary for describing classroom instruction benefits both teachers reflecting on their own instructional practices and researchers aiming to study instruction.

The possibility of decomposing quality of teaching into different components

Another important advantage of employing manuals for observing instruction is that they can aid researchers in systematically decomposing the quality of teaching into different components and thus identify critical factors. Observation systems embody a community of practice's view of high quality teaching and learning components) and will vary, emphasizing different decompositions of teaching and learning and implicitly referencing different goals of education and various roles for the teacher (Biesta and Stengel, 2016). These views also reflect cultural differences in social and cognitive values across communities and contexts (Paine et al., 2016; Stigler and Hiebert, 1999). In Japan, for example, an observation system might privilege measuring how well a lesson plan developed collaboratively with other teachers was implemented (Clarke et al., 2006) while observation systems in the Nordic countries might privilege student engagement and interaction. Communities' perspectives of teaching quality can be located along a continuum that moves from a behaviourist view of teaching and learning to a more cognitive view to a more sociocultural or situated view (Russ et al., 2016). Sociocultural approaches to teaching might emphasize interactive patterns, collaboration and students' opportunities to talk, while cognitive approaches might seek to tap into cognitive challenges, self-efficacy and the use of conceptions and misconceptions. Communities' perspectives, however, often blur the boundaries across this continuum and, depending on how thoroughly a system, or manual, is documented, it can be difficult to determine what view(s) underlie a specific perspective. Further, it is not helpful to dichotomize or essentialize views of instruction (Grossman and McDonald, 2008), as it can lead to a focus on differences in

how communities define and label teaching rather than a focus on how teaching and learning are related. In fact, there is some empirical evidence from meta analyses that suggests mixtures of dimensions of teaching that cut across these categories may be most important for student learning (Hattie, 2012; Gersten et al., 2009; Seidel and Shavelson, 2007).

While traditional coding schemes used in participant observation sharpen the analytical gaze of researchers and draw attention to specific decompositions of instruction, these possibilities are magnified through the combination of video data from classroom instruction and observation instruments. One of the reasons video data have become so popular in the educational sciences has to do with video being a real-time sequential medium (Jewitt, 2012). Recording instruction makes it possible to capture how temporal sequences unfold in the classroom and analyse them in detail. Several scholars have emphasized how the time-consuming nature of video data often results in only small segments of data being analysed and how this might lead to magnification of somewhat or potentially insignificant events (Blikstad-Balas, 2016; Lemke, 2007; Snell, 2011). With observation instruments and video data, the rigor of research might increase, because teaching can be decomposed into a variety of different components that can all be analysed in detail. For instance, the aforementioned PLATO instrument (Grossman et al., 2013) allows for systematic focus on the quality of 13 different elements of instruction. What components of teaching should be investigated and how will, of course, never be set in stone, and we do not wish for educational researchers to scrutinize only aspects of teaching that are well described in a manual. Yet, researchers who *are* attempting to scrutinize the same aspects of instruction may benefit from using the same way of decomposing that particular aspect of teaching.

Observation instruments allows us to compare quality of instruction across classrooms

Whether an existing instrument can or should be used or whether there is a need to develop a unique set of codes for a specific project always deserves careful consideration. That being said, we claim that with regard to studies of instructional quality, the field is in desperate need of comparable studies from different contexts. Lemke (2007: 35) has called attention to how educational research includes thousands of excellent analyses of 5-minute episodes from different classrooms but at the same time contains very few analyses of entire lessons and "almost none of either whole school days for individual pupils or teachers or a whole week (much less a whole year) in the life of one class". One problem with analysing short segments of video data with codes that are unique to a particular

study is that it is virtually impossible to relate them to other studies or compare findings with other research in a reliable way. By using existing observational instruments, it may become possible to compare findings from different studies across contexts (for example, across grades, subjects or countries) and by doing so, obtain more information about a given teaching practice across these different contexts. For instance, in an ongoing comparative classroom study (in progress), we compare classroom instruction in mathematics and language arts in Norway and Finland. By employing a validated coding manual and by using certified raters, we are able to make reliable and systematic comparisons between instructional practices in these two countries and thus discuss the differential impact of teachers' use of, for example, feedback practices, modelling and classroom discourse in Finnish and Norwegian mathematics classrooms. Common tools for research developed within an integrated methodological design can help researchers make progress in systematically describing possible differences in instructional practices across contexts as well as providing systematic knowledge about the impact of different teaching approaches across settings and subjects areas. This may provide important information not only about differences and similarities between teachers across contexts but it may also provide important knowledge about how different instructional approaches may have different impact on different students. What works in one context might not work in another, and by using validated observation instruments in combination with other data such as achievement data, copies of student work, assignments, etc., one can go beyond comparing, for instance, test results across nations. In an increasingly global educational discourse, it is of paramount importance that differences in achievement and instructional practices between countries be investigated not only with local instruments and small-scale studies but also with in-depth studies that are tailored for comparison of instruction. While there are robust international tests of student achievement (e.g. PISA and TIMSS), there is less tradition for rigorous international comparison of naturally occurring classroom instruction.

II A brief overview of coding manuals

In the following, we provide an overview of selected observation manuals with regards to (i) views of learning, (ii) content coverage, (iii) domains investigated, (iv) scoring systems and (v) certification requirement. We have chosen manuals that attempt to capture classroom instruction in one or more specific subjects. As summarized in Table 1, four manuals are included in our review.

CLASS (Classroom Assessing Scoring System)

The Classroom Assessing Scoring System (CLASS) is perhaps the most well-known of all observation manuals in educational research. It was developed at the Curry School's Center for Advanced Study of Teaching and Learning (CASTL). CLASS identifies multiple dimensions of interaction that have been linked to student achievement and student development. While CLASS was originally developed for an American study on early childhood development, it has been adapted to other school contexts and is used across a range of subjects. There are currently six different versions of CLASS adapted for observations in the following age groups: infant, toddler, pre-K, K-3, upper elementary, and secondary.

CLASS builds on a communicative approach to teaching and learning and interactions between students and adults are seen as the primary mechanism of student development and learning. In the CLASS manual, teaching is conceptualized through three broad domains of effective interactions: Emotional Support, Classroom Organization, and Instructional Support. Each of these domains further comprises multiple dimensions of effective interactions that have been proven to contribute to students' success in school, such as Teacher Sensitivity, Language Modelling, Behaviour Management, and Quality of Feedback. In total, CLASS covers ten different dimensions of instruction.

CLASS requires certification. A certified CLASS observer will conduct cycles of 15-minute observations of teachers and students in a given situation. Using the CLASS manual and its descriptions of behaviours and responses, the observer will then rate the ten CLASS dimensions of teacher–child interactions using a scale from 1 to 7.

CLASShas been used for observations in over six thousand classrooms, and several studies have found that higher CLASS ratings correlate with social gains and achievement gains. The CLASS instrument was one of four observation manuals included in the Bill and Melinda Gates–funded study of teaching, the Measures of Effective Teaching Project (hereafter, the MET project). A detailed overview of CLASS can be found at http://curry.virginia.edu/research/centers/castl/class.

FFT (Framework for Teaching)

The Framework for Teaching, developed by Charlotte Danielson (Danielson, 2013), identifies aspects of teaching that have been documented through research as promoting improved student learning. According to the Danielson group, the FFT can serve as the foundation of a school or

district's mentoring, coaching, professional development and teacher evaluation processes. The FFT aims to make explicit what good teaching is, and the observation instrument thus attempts to define what teachers should be able to do when exercising their profession. The FFT is generic and can be applied across school subjects and levels. The original framework was introduced in 1996 and was revised in 2007 and in 2011. The most recent revision of FFT was in 2013.

Danielson argues that the FFT framework builds on constructivist approaches to teaching and learning, underscoring student engagement and teacher–student interaction as key aspects of high quality teaching and learning. In the FFT, teaching is conceptualized through four overarching domains of teaching responsibility: Planning and Preparation (domain 1), Classroom Environment (domain 2), Instruction (domain 3), and Professional Responsibilities (domain 4). These four domains are further divided into 22 components and 76 smaller elements. The FFT is concerned with what happens in the classroom during lessons (domains 2 and 3) as well as with dimensions of teaching that are not directly observable in a given lesson, that is, domains 1 and 4 address planning and preparing lessons and professional responsibilities such as the professional thinking that occurs after an instructional event, communicating with families and showing professionalism in the community. In this framework, teachers are evaluated through a rubric with a score from 1 to 4. They can be ranked or measured as either *unsatisfactory*, *basic* (low end), *proficient*, or *distinguished* (high end). All categories are explicitly described in the rubric, and critical attributes of each category are provided and examples are given.

The following rubric for unsatisfactory (low end) versus distinguished (high end) scores from the domain Instruction (domain 3), component "Communicating with students" (3a), serves as an illustration:

Unsatisfactory	Distinguished
The instructional purpose of the lesson is unclear to students, and the directions and procedures are confusing. The teacher's explanation of the content contains major errors. The teacher's spoken or written language contains errors of grammar or syntax. The teacher's vocabulary is inappropriate, vague, or used incorrectly, leaving students	The teacher links the instructional purpose of the lesson to student interests; the directions and procedures are clear and anticipate possible student misunderstanding. The teacher's explanation of content is thorough and clear, developing conceptual understanding through artful scaffolding and connecting with students' interests. Students contribute to extending the content

confused.	and help explain concepts to their classmates. The teacher's spoken and written language is expressive, and the teacher finds opportunities to extend students' vocabularies.

A pdf version of the FFT manual (including the rubrics) is available from the Danielson group website; there are no certification or training requirements.

The FFT is a validated observation instrument and was one of the observation instruments used in the previously mentioned MET study. It is also widely used in teacher evaluations, and the 2013 edition specifically identifies instruction that promotes student learning in the context of the American Common Core State Standards. For a detailed account of FFT, see the Danielson group website: https://www.danielsongroup.org/framework/.

The PISA+ manual

The PISA+ manual (Klette et al., 2005) was developed to analyse teachers' instructional practices in the three PISA (Programme for International Student Assessment) domains mathematics, science and reading in lower secondary classrooms in Norway. The instrument combines process–product and sociocultural approaches to learning and was developed as a subject-generic instrument; however also supported with more subject-specific instruments (see Ødegaard and Arnesen, 2007). targeted to capture specific aspects such as discourse features and scientific talk in science classrooms for example. Drawing on empirical research from classroom teaching and learning and existing coding manuals (such as the CLASS manual), the PISA+ manual distinguishes between Instructional Formats, Physical Organization (environment) and Time Management as the three main domains. Instructional formats are divided into whole-class instruction, individual seatwork and group work/work in pairs as subareas, with subcategories linked to each area. For example, whole-class instruction is divided into monologic instruction, dialogic instruction and question–answer sequences as three different forms of teacher-led whole-class instruction. In addition, the category of whole-class instruction captures whole-class discussions (defined as students commenting on each other without teacher interventions), student presentations and task management. Raters are trained to a satisfactory reliability level (70% inter-reliability level); however, there is no formal certification requirement linked to this manual, nor any formal training facilities.

Contrary to many of the other coding manuals, the PISA+ manual uses frequency coding, and thus coding takes place whenever the activity appears throughout the class period, using the whole lesson (e.g. 45 minutes, 60 minutes, 70 minutes) as the analytical unit. There are no scoring rubrics linked to the PISA+ manual and the categories basically capture the *what* of the activity (present–not present) rather than *how* the activity is deployed [e.g. the distinction between low end (*unsatisfactory*, *basic)* and high end (*proficient*, *distinguished*) proficiency as in other coding manuals].

Although the PISA+ manual could serve as an example of a "local instrument", it shares features with other manuals such as the PLATO manual and the CLASS manual and has been used in comparative classroom research (Ødegaard and Arnesen, 2007), researcher training (Dalland, 2011) and evaluating school reform efforts (Klette, 2015). Further, the systematic use of this manual has paved the ground for increased interest in methodological aspects of studies of classroom teaching and learning, including how instruments, procedures for coding and available data and access might support joint analyses and strengthen the (re)use of existing data (Andersson and Sørvik, 2013; Hammersley 2010).

PLATO (Protocol for Language Arts Teaching Observation)

The Protocol for Language Arts Teaching Observations (PLATO) is a classroom observation protocol developed by Pam Grossman at Stanford University that aims to capture features of English/Language Arts (ELA) instruction. The protocol was originally developed to investigate the relationship between teachers' classroom practices and their impact on student achievement. It is currently used as a professional development tool to promote teachers' use of rigorous, research-based teaching practices. PLATO is designed to work across a variety of curricula and instructional approaches (within ELA); the current version is PLATO 5.0.

The PLATO instrument is organized around four instructional domains: Instructional Scaffolding, Disciplinary Demand, Representing and Use of Content and Classroom Environment. PLATO domains are further categorized as 13 sub-elements of instruction, for example Modelling, Intellectual Challenge, Text-Based Instruction and Time Management.

PLATO requires certification. Each of the 13 elements are scored on a scale from 1 to 4 based on the evidence for a given element (for example, Modelling) during a 15-minute cycle. At the low end (scores 1 and 2), there is almost no evidence or little evidence of instructional practice related to the

element in question, while the higher end (scores 3 and 4) is characterized by evidence with some weaknesses or strong and consistent evidence, respectively. In addition to the 13 elements, PLATO captures the content of instruction (for instance writing, literature and grammar) as well as the overall activity structures (whole group, small group, independent work, etc.) for each 15-minute segment.

Several studies have identified significant relationships between PLATO scores and gains in student achievement (Grossman, Cohen, Ronfeldt and Brown, 2014; Grossman, Cohen and Brown, 2014; Kane and Staiger, 2012).

For a detailed account of PLATO, see Grossman, Loeb, Cohen, and Wyckoff (2013) as well as the website: http://platorubric.stanford.edu/index.html.

In Table 1 below, we summarise features of the reviewed coding manuals.

INSERT TABLE 1 about here

Table 1. Summarised overview of the coding manuals

III Pitfalls and possibilities of observation manuals

In the following section, we will briefly compare some of the presented manuals in order to discuss the possibilities and challenges of using systematic observation instruments.

The manuals we have reviewed vary in subject specificity. PLATO is particularly relevant for language arts teaching and secondary level, while CLASS was developed for all subjects and different versions are adapted to different levels in the school system. Thus, CLASS, FFT and the PISA+ manual can all be said to be generic, while PLATO was tailored for a specific subject (although it is currently being used in other subjects as well). Another important difference is that the FFT has a wider scope than the other manuals; it is not limited to classroom instruction happening in an observable lesson. Further, the specific focus on the Common Core State Standards and the many local examples (for example, specific American organizations and local holidays) makes the FFT less adaptable to other contexts, and thus less relevant for international comparisons. Another difference between the protocols is their accessibility for the research community. To become certified in

CLASS, one may order a training session or become certified through an online course[1]. Becoming certified in PLATO requires collaboration with and certification by the developers of PLATO at Stanford University (online version available), while those who wish to use the PISA+ framework or the FFT have free access and opportunity to use the tools.

In examining features considered to be generic elements of teaching, the reviewed protocols share a commitment to investigating aspects of classroom learning such as instructional format, instructional clarity, cognitive challenge, how teachers and student interact (discourse features) and classroom environment. Despite slightly different approaches to learning, they all highlight student engagement, instructional clarity and classroom climate as key components when analysing features of classroom learning. Across the different manuals there seems to be a shared understanding of key elements of qualities in teaching (e.g. instructional clarity, student engagement and classroom climate) that could be investigated, measured and developed, thus helping researchers and practitioners to analyse "what works" and for what purpose across different classroom settings and subjects. These manuals all offer examples of views of teaching and learning that support high levels of student participation, value the knowledge that students bring to the classroom and maintain high levels of academic rigor. In this sense, these manuals suggest a descriptive and empirical — and also a normative — take on classroom teaching and learning.

The previously mentioned MET study used five different classroom observation manuals [CLASS, FFT, PLATO, the Mathematical Quality of Instruction (MQI) and the Quality Science Teaching (QST)] in an attempt to measure effective teaching fairly and reliably. They found that (i) well-designed manuals could provide reliable feedback on aspects of teaching that are predictive for students learning; (ii) accurate ratings of teaching require two or more lessons, each scored by a different certified rater; and (iii) estimates are more stable if observation manuals are combined with other measures such as student surveys and/or achievement scores (BMGF, 2012). Last but not least, they found that subject-generic measures such as CLASS and FFT had greater reliability as estimates for student learning (>60%), while subject-targeted manuals such as PLATO and MQI showed lower reliability (<60%). However, more recent analyses (Hill and Grossman, 2013) have shown that the variation in reliability among these measures is less distinct than set out in the early reports and that there are strong similarities in predictive values between the subject-generic manuals

[1] http://teachstone.com/class-trainings/class-observation-training-programs/

FFT and CLASS and the subject-specific manual PLATO (60%), with a slightly lower value for the case of the subject-specific manual MQI.

While we have highlighted many of the benefits of drawing on observation manuals, we would also like to emphasize the many drawbacks and caveats. A key challenge and critique of standardized coding schemes such as the ones reviewed here is that they are reductionist (Blikstad-Balas, 2016; Snell, 2011). No manual can ever capture everything going on in a social setting, and it is in the very nature of codifications to reduce. In the manuals we have reviewed, the complexity of teaching is reduced to a set of decontextualized codes. In this way, different teachers with very different students and different pedagogical approaches will be "measured" by the same normative standards. In some cases, this might be problematic. For example, if a given manual is tailored to one specific context (e.g. "mathematic lessons"; kindergarten level; American classrooms), there might be a systematic bias if researchers are comparing different groups across levels and nations. As we already commented, we believe the fact that the close linkage of the Framework for Teaching to an American context and the CCSTs makes it less relevant for other contexts without adaptation. While PLATO has been successfully used across contexts and subjects, it would be unfair to expect, for instance, a category like "text-based instruction" to score as highly in mathematics as in language arts.

Another challenge is that coding is closely related to magnification (Blikstad-Balas, 2016; Lemke, 2007; Snell, 2011). Relying on pre-defined and pre-validated codes may result in an overview of rich data that, if not scrutinized further, may be a misleading magnification. For example, in a study conducted by Brevik, Klette & Blikstad-Balas (2016), the use of PLATO showed that a large number of teachers score low (2) on the category Feedback. While such coding is very useful because it enables more consistent and more comparable data, it is also important to look into variations within one applied code. In this specific case, it would be misleading to assume that all the analysed lessons scoring low in a given category (such as Feedback) are a result of very similar instruction — this needs to be studied in context, and in detail. In fact, a qualitative analyses of the teachers that scored high on the category Feedback (Dåsvatn, 2016) showed that teachers who all scored on the high end (4) in the category Feedback had different instructional practices. Thus, while these teachers have in common that they are all able to provide their students with high quality feedback, their ways of doing so differ significantly. To avoid exaggerated magnification, then, we must not

overlook the similarities and the differences in instruction within each code, which will highlight what these lessons may have in common beyond the common code they have been assigned.

More recently, systematic use of coding manuals has been combined with student achievement scores for the purpose of distinguishing how the different instructional practices (e.g. discourse features, feedback, instructional clarity) align with students' learning. One finding across these studies is that classroom climate (time and behavioural management) is favourable for student learning (Doyle, 1986; Scheerens, 2014); it has been harder to establish such a link for instructional practices such as discourse features and cognitive challenge. While this is an important finding, it also touches upon a key challenge of observation manuals in general, as it may reflect that measurement of time management and behavioural management is easier and more reliably performed than more abstract, high inference features. Further, these two categories (time and behaviour management) are measures that were developed in the early 1970s and have undergone refinement since, while the systematic and codified measurement of discourse features, for example, is still is in its infancy, suffering from less reliability and rigour. Our point is that coding manuals often combine newer, high-inference elements such as discourse features or intellectual challenge with classical lower inference elements such as time management. If only the latter have predictive power when it comes to student achievement, this may not only suggest the importance of time management but also point to a difficulty in measuring for example discourse features in a precise and reliable way.

What is often referred to as the law of the instrument suggests that "if all you have is a hammer, everything looks like a nail". Observation manuals can be perceived as lenses that guide the search for specific features, thereby always neglecting others. If you are a certified scorer trained to master a specific manual, you are "wearing the lenses" and views of learning of that manual, and it is only natural that the majority of your attention is on the specific codes you are assigning to the data. This may be both the major benefit and the major drawback of using a set of predefined codes. While observation manuals sharpen the analytical gaze by explicitly defining what to look for and how to recognize and assess it if we find it, they also limit the same analytical gaze by systematically steering our attention to what the manual considers relevant.

A final challenge with observation manuals that deserves attention is that they can contribute to research that merely reproduces itself if the same manuals are used to address similar questions again

and again Although we have advocated that manuals such as the ones reviewed in this article offer a systematic lens through which to view classroom teaching, it is also of paramount importance to consider in each case whether they measure what we need to know more about and not just what we have good measures for.

Concluding Remarks

Observation manuals have been a part of our methodological toolkit for decades. However, recent developments in video technology and available software programs have turned this approach to analysing teaching and learning into a powerful means of rigorous analysis of teaching and learning practices in classrooms as well as a tool for teacher education and professional development. Video data, together with new annotation tools, make it possible to record, review, edit, analyse and synthesize different instructional practices, which is beneficial for researchers as well as for teachers and teacher educators. Supported with systematically developed and validated coding manuals, video documentation represents what we might describe as a breakthrough in research on classroom teaching and learning, giving us the opportunity to test out the differential power, implications and possible practices linked to different instructional practices. As such, the development and use of common observation manuals could be one way, out of several, to establish a framework for comparative didactics while at the same time starting the meticulous task of establishing a common language and vocabulary for analysing and describing teaching.

We have argued that coding manuals might strengthen conceptualizations and opportunities for studies of classroom teaching and learning for teachers as well as for researchers. What becomes important is not a consensus on a final set of universal teaching practices but rather a continual dialogue within the field and among scholars on how to conceptualize aspects of practice that support practitioner learning of high quality instruction. In this sense, observation manuals may contribute to a shared and explicit understanding of teaching because they are an attempt to verbalize what the craft of teaching *is* and how it can be identified through systematic observation of what teachers do.

In his classical study, dating back to 1975, Dan Lortie (1975) describes the cellular structure of the American education system, describing teachers mainly as lonely professionals, working in their individual classrooms, in schools organized in an egg carton format and lacking a shared technical language and vocabulary. As argued throughout this article, coding manuals and video

documentation could be one way out of the longstanding legacy of "tinkering" (Hargreaves, 2000) and idiosyncrasy (Ball and Cohen, 1999) as the main mechanisms of generating knowledge in studies of classroom teaching and learning.

References

Alexander, R. J. (2001): Culture and Pedagogy: international comparisons in primary education. Oxford and Boston: Blackwell.

Andersson E and Sørvik GO (2013) Reality lost? Re-use of qualitative data in classroom video studies. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research* 14(3): Art. 1.

Ball DL and Cohen DK (1999) Developing practice, developing practitioners: Toward a practice-based theory of professional education. In: Sykes G and Darling-Hammond L (eds) *Teaching as the Learning Profession. Handbook of Policy and Practice.* San Francisco, CA: Jossey Bass, pp.3–32.

Biesta, G. J. J., & Stengel, B. (2016). Thinking philosophically about teaching: Illuminating issues and (re)framing research. In D. Gittomer & C. Bell (Eds.), AREA handbook of research on teaching (5th ed.). Washington, DC: AERA.

Blikstad-Balas, M. (2014). Vague concepts in the educational sciences: implications for researchers. Scandinavian Journal of Educational Research, 58(5), 528-539.

Blikstad-Balas, M. (2016). Key challenges of using video when investigating social practices in education: contextualization, magnification, and representation. International Journal of Research & Method in Education, (ahead-of-print)1-13.

Blikstad - Balas, M., & Sørvik, G. O. (2015). Researching literacy in context: using video analysis to explore school literacies. Literacy, 49(3), 140-148.

BMFG: Bill and Melinda Gates Foundation (2012). Gathering feedback for teaching: Combining high quality observations with student surveys and achievement gains. Seattle, WA: The Bill and Melinda Gates Foundation.

Brevik, L., Klette, K. & Blikstad-Balas, M. (2016) *The Quality of Feedback: Instructional Practices Captured in Video-Recorded Classroom Observations.* Paper presentation at AERA conference, april 12. 2016; Washington DC

Cohen, D., & Moffitt, S. (2009). *The ordeal of equality. Did federal regulation fix the schools?* Cambridge, MA: Harvard University Press.

Clarke, D., Emanuelsson, J., Jablonka, E. & Chee Mok, I. A. (Eds.), (2006a). *Making Connections: Comparing Mathematics Classrooms Around The World.* Rotterdam: Sense Publishers.

Dalland C (2011) Challenges when using qualitative data gathered by others (Ufordringer ved gjenbruk av andres kvalitative data). *Norwegain Journal of Eductaion/ Norsk Pedagogisk Tidsskrift* 6: 449–459).

Danielson C (2013) *The Framework for Teaching Evaluation Instrument* (2013 Edition). Danielson Group. Available at: www.danielsongroup.org (accessed 3 July 2016).

Dreeben, R. (2005). Teaching and the competence of occupations. In L. Hedges & B. Schneider (Eds.),*The social organization.*

Doyle W (1986) Classroom organization and management. In: Wittrock MC (ed) *Handbook of Research on Teaching. A Project of the American Educational Research Association*. New York, NY: Macmillan, pp.392–431.

Dåsvatn M (2016) Muntlige tilbakemeldinger i norskfaget. En videostudie av fem klasserom. [Oral feedback in Norwegian language arts. A video study of five classrooms.] Master thesis, University of Oslo.

Fischer H and Neumann K (2012) Video analysis as a tool for understanding science instruction. In: Dillon and Jorde D (eds) *The World of Science Education.* Rotterdam, Netherlands: Sense Publishers, pp.115–140.

Foray D. (2001) Facing the problem of unbalanced development of knowledge across sectors and fields: The case of the knowledge base in primary education. *Research Policy* 30: 1553–1561.

Foray D. (2010) Educational innovation – An economist's perspective. In: Author 1 (ed) *Rigor and Relevance in Educational Research. Report from the Mars Seminar, 2010*. Oslo, Norway: The Research Council of Norway, pp.35–45.

Foray D. and Hargreaves D. (2003) The production of knowledge in different sectors: A model and some hypotheses. *London Review of Education* 1: 7–19.

Gersten, R., Chard, D. J., Jayanthi, M., Baker, S. K., Morphy, P., and Flojo, J. (2009). Mathematics Instruction for Students with Learning Disabilities: A Meta-Analysis ofInstructional Components. Review of Educational Research, Vol. 79, No. 3, pp. 1202-1242.

Groninger, R. G., Valli L. and Chambliss M.J. (2012) Researching quality in teaching: Enduring and emerging challenges. *Teachers College Record* 114: 1–16.

Grossman, P and MacDonald, M. (2008) Back to the future: Directions for research in teaching and teacher education. *American Education Research Journal* 45(1): 184–205.

Grossman P., Loeb S., Cohen J. and Wyckoff J. (2013) Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education* 119(3): 445–470.

Grossman P., Cohen J., Ronfeldt M. and Brown L. (2014) The test matters: The relationship between classroom observation scores and teacher value added on multiple types of assessment. *Educational Researcher* 43: 293–303.

Grossman P., Cohen J. and Brown L. (2014) Understanding instructional quality in English Language Arts: Variations in the relationship between PLATO and value-added by content and context. In:

Hammersley M. (2007) The issue of quality in qualitative research. *International Journal of Research & Method in Education* 30(3): 287–305.

Hammersley M. (2010) Can we re-use qualitative data via secondary analysis? Notes on some terminological and substantive issues. *Sociological Research Online* 15(1): 5.

Hargreaves D. (2000) The production and mediation of knowledge among teachers and doctors. A comparative perspective. In *OECD (2000): Knowledge Management in the Learning Society*. Paris, France: OECD, pp.219–238.

Hattie, J. (2012): Visible Learning for Teachers: Maximizing Impact on Learning. Milton Park: Routledge.

Hill, H., & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review*, *83*(2), 371-384.

Hudson B. and Meyer M. A. (eds) (2011) *Beyond Fragmentation: Didactics, Learning and Teaching in Europe*. Budrich, pp.9–28.

Jewitt C. (2012) An introduction to using video for research. NCRM Working Paper. National Center for Research Methods. Available at: http://eprints.ncrm.ac.uk/2259/ (Accessed 3 July 2016).

Kane T., Kerr K. and Pianta R. (eds) *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project*. John Wiley & Sons.

Kane, T. J., & Staiger, D. O. (2012). *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains*. Research Paper. MET Project. Bill & Melinda Gates Foundation.

Klette, K. (2009). Challenges in strategies for complexity reduction in video studies. Experiences from the PISAC study. In T. Janik & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 61–83). Münster: Waxmann Publishing.

Klette, K. (2010). Blindness to change during processes of change: What do educational researchers learn from classroom studies? In A. Hargreaves, A. Libermann, & M. Fullan (Eds.), Second international handbook of educational change. Amsterdam: Springer Publishing.

Klette, K (2015). Introduction: Studying Interaction and Instructional Patterns in Classrooms. In, Klette, K, Bergem, OK, Roe, A. (eds): Teaching and Learning in Lower Secondary Schools in the Era of PISA and TIMSS. (pp. 1-14) New York : Springer

Klette, K., Lie, S., Anmarkrud, Ø., Arnesen, NE; Bergem, OK, Ødegaard, M., & Zachariassen, JRH. (2005). Categories for video analysis of classroom activites with a focus on the teacher. URL: http://www.uv.uio.no/ils/english/research/projects/pisa-pluss/coding/Categories%20for%20video%20analysis%20of%20classroom%20activities%20with%20focus%20on%20the%20teacher.pdf

Kuger, S., & Klieme, E. (2016). *Dimensions of context assessment. In Assessing Contexts of Learning* (pp. 3-37). Springer International Publishing.

Lortie D. (1975) *School Teacher. A sociological perspective*. Chicago, IL: Chicago University Press.
Lemke J (2007) Video epistemology in-and-outside the box: Traversing attentional spaces. In:

Maxwell J.A. and Chmiel M (2014) Notes toward a theory of qualitative data analysis. In: Flick U (ed) *The SAGE Handbook of Qualitative Data Analysis*. Sage, pp.21–35.

Morse, J.M. (2015). Critical Analysis of Strategies for Determining Rigor in Qulitative Inquiry. Qualitative Health Research 29 (9), pp 1212-1222.

Nelson R. (2000) *Knowledge and Innovation Systems. In OECD Knowledge Management in the Learning Society (CERI) report (pp 115-125)*, Paris, France: OECD.

Paine, L., Bloemeke, S., & Aydarova, O. (2016). Teachers and teaching in the context of globalization. In Gitomer, D. and Bell, C. (eds)*Handbook of research on teaching*, 717-786. Washington, DC : American Educational Research Association

Pianta R.C., La Paro K. and Hamre B.K. (2008) *Classroom Assessment Scoring System (CLASS).* Baltimore, MD: Paul H Brookes.

Rich P.J. and Hannafin M. (2009) Video annotation tools technologies to scaffold, structure, and transform teacher reflection. *Journal of Teacher Education* 60(1): 52–67.

Russ, R.., Sherin, B.L. and Sherin, M.G. (2016) What Constitutes Teacher Learning? In Gitomer, D. and Bell, C. (eds) *Handbook of research on teaching*,. 391-438 Washington, DC : American Educational Research Association

Saldaña J. (2013) *The Coding Manual for Qualitative Researchers* (2nd ed). Thousand Oaks, California: Sage.

Scheerens J. (2014) School, teaching, and system effectiveness: Some comments on three state-of-the-art reviews. *School Effectiveness and School Improvement* 25(2): 282–290.

Seidel T.,Shavelson R. J. (2007). Teaching Effectiveness Research in the PastDecade: The Role of Theory and Research Design in Disentangling Meta Analysis Results. *Review of Educational Research, Vol 77*(4), 454-499.

Snell J. (2011) Interrogating video data: Systematic quantitative analysis versus micro-ethnographic analysis. *International Journal of Social Research Methodology* 14(3): 253–258.

Stigler J. & Hiebert (1999). The teaching gap: Best ideas from the world's teachers for improving education in the classroom. New York: Free Press.

Stuhlman M.W., Hamre B.K., Downer J.T. and Pianta R.C. (2010) Why should we use classroom observation?. Centre for Advanced of Teaching and Learning, University of Virginia. Available at: http://curry.virginia.edu/uploads/resourceLibrary/CASTL_practioner_Part1_single.pdf (accessed 23 August 2014).

Sun J. and van Es E.A. (2015) An exploratory study of the influence that analyzing teaching has on preservice teachers' classroom practice. *Journal of Teacher Education* 66(3): 201–214.

Tyack D., and Cuban L., (1995). *Tinkering toward utopia. A Century of Public School Reform*. Cambridge, MA: Harvard University Press.

Ødegaard, M. & Arnesen, N.E. (2006). Categories for video analysis of science classroom activities. Oslo: University of Oslo.