Global Teaching InSights

Technical Report

Section II: Instrument development

4 The development of the Study observation coding system

Courtney A. Bell

Specific high-level design principles guided iterative rounds of observation protocol development to ensure that the Study's observation coding system could be meaningfully applied across participating countries/economies. This chapter describes the nature of the intended claims as well as the design principles used to create evidence that could then support the claims. In what follows, the chapter elaborates the development process for the two types of observation codes – components and indicators – that were used to rate the video recorded lessons in the Study.

Introduction

This chapter describes the process of developing the Global Teaching InSights (results from the TALIS Video Study project, and is hereafter cited in this chapter as "the Study" or "GTI") observation system using evidence-centred design (ECD) assessment principles and the shared conceptualisation of teaching quality developed during the review and harmonisation activities described in Chapters 1 and 2. Using an iterative, collaborative code development process, two types of codes - component and indicator – were used to quantify teaching quality through video observations.

The words and phrases that comprise the observation codes refer to complex socially-situated classroom behaviours that have varied meanings within and across countries/economies (Paine, Bloemeke and Aydarova, 2016[1]). Some classroom behaviours are clearly visible, such as when students work in groups or laugh with one another. But others are harder to perceive because they are subtle, occur in the context of other behaviours, and evolve over multiple interactions between teachers and students. For example, a student may show confusion by raising her hand and asking a question or she may show confusion by waiting until the teacher passes by the students' desk during the teachers' routine circulation around the room. The observation codes focus attention on a variety of teaching practices that differ in the degree to which observers can "see" the behaviours that comprise them.

The Study's observation codes provide one common language to refer to teaching. The coding protocols were designed to capture aspects of teaching that transcend national boundaries, but to also be sensitive to the variation that can exist between and within cultures and countries. The observation codes do this by focusing on agreed to observable evidence that allows for inferences about the type, amount, and levels of specific teaching practices.

Developing the Study's observation coding system

Reflecting a community's view of teaching quality

Any new assessment instrument should reflect a community's view of that domain or construct. Chapters 1 and 2 describe the review and harmonisation activities carried out by the International Consortium. These activities harmonised international research on teaching quality, international frameworks (e.g. those of PISA and TALIS), and participating countries'/economies' standards of teaching quality. The observation code development team (hereafter the observation team) used the resultant six domains of teaching quality (see Table 2.3 in Chapter 2) as the domains to be operationalised in the observation codes.

Intended claims

In order to determine how to operationalise the six domains into codes that raters could use, the observation team followed an evidence-centred assessment design (ECD) approach to start developing the Study's observation coding system. ECD is defined as an assessment framework "that makes explicit the structures of assessment arguments, the elements and processes through which they are represented, and the interrelationships among them" (Mislevy, Steinberg and Almond, 2003, p. 6[2]). The observation team's first step was to determine the types of claims the observation codes needed to support.

The observation scores needed to support the following claim: in the observed lesson there was a [level, nature or type] of a specific teaching practice. For example, one such claim might be "In lesson A, there was a moderate level of subject matter quality present from which students could learn." It is important to note that the foundational claim was not that certain practices were better or more valuable than other practices, but rather that scores conveyed the quality, nature or presence of a specific practice. Only in the later analytic stage when specific practices were analytically related to desired student outcomes could those types of inferences be determined.

Design principles

The claim that codes specified the quality, nature or presence of a specific practice required five design principles. Codes were collaboratively developed to reflect global conceptions of teaching quality and to capture variation in teaching within the eight participating countries/economies. The codes also had to be scalable so that they could be used in a train-the-trainer model of implementation. This meant the codes had to be defined and operationalised in standardised ways so that bilingual raters could apply them. And finally, to support standardised application across countries/economies, codes had to focus on behaviours that are observable on video. Described in more detail below, these five design principles guided the development of the Study's observation codes.

- **Collaborative code development**. Research has long criticised "drop from the sky" assessments and the ways in which they can be inappropriate – in what they measure, how constructs or practices are measured, and how scores are used (Linn, 1994[3]; Shepard, 1989[4]). For the codes to be valid and useful across countries/economies, a collaborative approach was used to specify what would be measured and how. Collaboration was necessary from the very first development tasks (defining teaching quality) to the very last tasks (finalising the training and quality control materials) (see the description of how the collaboration occurred in the next section).
- **Capture variation within and between countries/economies**. The observation codes needed to capture variation in teaching practice within and across the classrooms in all participating countries and economies. Repeatedly testing the codes on the pilot classroom videos helped focus development efforts around this goal. Nuances and discrepancies were discussed iteratively with participating countries and economies until there were a set of constructs that captured teaching adequately.
- **Scalable codes and training materials**. In contrast to studies in which the codes are never intended to be used with large numbers of classrooms or are only ever going to be used by a single research team with access to comparably affordable raters, the Study required a scalable and relatively affordable rating process. Codes and training materials were used across multiple countries/economies, with country experts (hereafter global master raters) leading within-country/economy rating activities. Ratings also needed to be created by comparably expensive bilingual raters. Rater training followed a train-the-trainer model and under this model, global master raters were trained by the observation team, who then trained and monitored bilingual raters in their respective countries/economies. Global master raters trained country/economy raters using the same English training materials with which they were trained and certified.
- **Standardised meanings across bilingual raters.** Threats to validity must be understood and managed in both national and international studies. Instrument reliability and rater agreement are particularly worrisome when human raters are involved in creating scores (Bell et al., 2015[5]; Casabianca, Lockwood and McCaffrey, 2015[6]; Floman et al., 2017[7]). To minimise these threats, the codes, training materials, processes and quality control procedures needed to be standardised.
- **Suitable to video coding.** The final design principle focused on the suitability of observations for capturing the valued teaching practice. Observation codes were developed for behaviours that are observable during lessons and about which raters can make inferences without significant additional information from other sources (e.g. an interview with the teacher or the entire quadratic equations unit plan). Observations of teaching give us reasonable access to behaviours and words, but poor access to the meaning or intention behind those behaviours and words. To standardise the meaning of codes across countries/economies, the Study's observation codes necessarily focused on behaviours.

These design principles meant that when a code was codifying subtle or culturally embedded nuances in classroom interactions that were difficult to discern in teacher and student behaviours, those codes were generally revised or dropped. If specific practices required the rater to have subtle or culturally specific knowledge, those practices were refined so that all countries/economies' raters could agree on them or the practices were not measured through observation codes. This principle also meant that certain aspects of teaching were not included in the observation codes – e.g. goals of the lesson, logical sequencing of activities and teachers' beliefs about student learning. To the degree possible these were captured in other study instruments.

Iterative code development

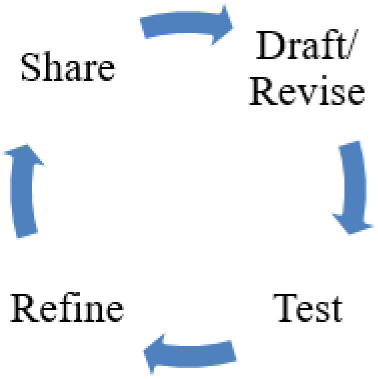
The development process for the two types of observation codes consisted of four cycles of drafting, testing, reviewing, and expert reviewing. This lasted more than two years. All development cycles used videos supplied by participating countries and economies to test codes against.

The design requirements were met by using multi-stage iterative development cycles guided by ECD and led by the multi-lingual, multi-cultural observation team. The list of global master raters and country/economy experts that participated in code development are listed in Annex F.

Based on the review and harmonisation activities described previously as well as discussions with the technical advisory group and the observation team drafted a set of constructs that were universally deemed important and were likely to be able to be turned into codes that could adhere to the design requirements.

Four cycles of iterative development over roughly two years of development activities resulted in five versions of the codes, the observation team drafted and tested the codes, refined them, then shared the codes with external experts and countries/economies, revised and then retested the codes and refined them further. Figure 4.1 shows the summary cyclical process.

Figure 4.1. Observation code development cycle



Source: OECD, Global Teaching InSights Database.

The four cycles incorporated both technical experts' and country/economy representatives' views early and often, beginning with the initial set (version 1) of codes. That increased the likelihood of the Study producing standardised, scalable codes and training materials that could be used by bilingual country/economy raters who were taught in a train-the-trainer model.

Each development cycle began with either the drafting or revision of the codes (draft/revise in Figure 4.1). Once the first set of draft constructs and codes were written, the observation team watched videos from all eight participating countries/economies and the United States, which participated in the initial phase of the Study to test and refine the codes. Each week, the observation team selected two videos from the pool of example videos collected from countries/economies. Members of the observation team watched the videos and tested the codes as they were written, rated each video, provided evidence for those ratings and, where applicable, noted how the codes were functioning. The entire observation team then met to share and discuss what was seen in the videos, agree on the ratings and discuss edits to the codes. After revising the codes, the observation team then selected two more videos to watch and test code for the following week's team meeting.

This process of testing and refining the codes continued for several weeks before the codes were shared with mathematics and observation experts from each country/economy for their feedback. Soliciting this feedback aimed to understand the degree to which teaching constructs were salient in each country's/economy's teaching and research communities, the level of variation that country experts expected to see of the teaching constructs in their classrooms, and to begin building a shared understanding with country/economy representatives about the teaching practices in the observation protocol. Countries'/economies' input was solicited in the written form and in face-to-face meetings. In-person meetings allowed the observation team to share short video clips with country/economy experts and ask them to test out the codes much like the observation team had done internally. When country/economy experts raised concerns pertaining to the degree of visibility of certain teaching practices in a particular country/economy, questions about how raters would find evidence to rate particular behaviours and suggestions for including additional components of teaching practice that were not captured by the existing codes, the observation team revised the codes and a new development cycle began again.

The group of country/economy experts played an active role in code development. They provided input on:

- Developing a shared understanding of the prevalence of codified behaviours in their classrooms and whether those behaviours meant the same thing across their education systems. The contextual factors relevant in their classrooms were used to redefine how the codes were specifying, capturing, and measuring the teaching constructs of interest.
- Constructs that seemed difficult to train a global community of raters to recognise in a standardised, reliable way. These were set aside and discussed with participating countries and economies at face-to-face meetings. Examples of such constructs include the pacing of the lesson and a teacher's expectations for student learning.
- Clarifying the markers of the teaching practices being measured. This included how raters would be able to find evidence to rate particular practices, and consensus on the types of behaviours that should "count" for specific constructs (e.g. deciding laughter should be counted as evidence of shared warmth).
- Suggestions for additional components of teaching practice that were not captured by the existing codes but that should be considered. This was important so that the codes captured the range of teaching practices present in all participating countries and economies.

Frequently while iterating on the codes, the observation team was concerned with how a code would apply in a specific country/economy (or specific lesson) and how specific constructs were being defined vis-à-vis the codes. The process of testing the codes and seeking feedback helped address the first concern, ensuring that the codes captured the range of teaching practices present in all participating countries/economies. For example, the classroom management techniques used in British classrooms differed from those found in Japanese classrooms or in Latin-American classrooms. The observation instruments needed to capture the variation in teaching practice within and across the classrooms in all countries/economies, and repeatedly testing the codes on videos from the participating countries/economies helped focus development efforts around this goal.

Due to the iterative nature of the development process, the codes and training materials had never been empirically tested until the main study. Thus, Global Teaching InSights (resulting from the TALIS Video Study project) was the first empirical test of the codes.

As Table 4.1 shows, there were five versions of the codes (numbered 0 – 4) developed over roughly two years of development activities.

Table 4.1. Observation code development process

Version/development cycle step	1st quarter of 2016	2nd quarter of 2016	3rd quarter of 2016	4th quarter of 2016	1st quarter of 2017	2nd quarter of 2017	3rd quarter of 2017	4th quarter of 2017	1st quarter of 2018	2nd quarter of 2018
0/drafted	x	x								
0/tested and refined		x	x							
0/shared			x							
0/revised			x	x						
1/tested and refined				x						
1/shared				x						
1/revised					x					
2/tested and refined					x	x				
2/shared						x				
2/revised						x				
3/tested and refined							x			
3/shared								x		
3/revised									x	
4/tested and refined									x	
4/shared										x

Note: The observation team shared internally tested and refined codes during in-person meetings with country/economy representatives in June 2016, November 2016, October 2017 and May 2018; they solicited written feedback in April 2017.
Source: OECD, Global Teaching InSights Database.

The Study's observation coding system

The observation coding system reflected on the six domains of quality teaching (see Table 2.3 in Chapter 2). First, the section describes how the observation coding system is defined and it provides details about the observation system – both the component and indicator codes. Second, it provides details about the characteristics of the Study's observation coding system.

Observation systems

Whenever an observation rubric or checklist is applied to a lesson, it is applied in an observation coding system (Hill, Charalambous and Kraft, 2012[8]). Observation coding systems are comprised of rating specifications, rating processes, and sampling and scoring specifications (Bell et al., 2019[9]; Liu et al., 2019[10]). The following paragraphs explain each one below and then describe in detail the Study's rating specifications. The rating process and sampling and scoring specifications are described in Chapter 6.

Rating specifications. Aspects of teaching, scales and standards in master ratings are the tools raters use to assign their ratings; essentially, they are scoring tools (Table 4.2). Aspects of teaching are the components or indicators that codify teaching (e.g. eliciting student thinking, persistence, and accuracy). Scales are the categories raters use to describe the nature, quality, presence, or frequency of the scales. For example, clarity may be rated on a scale from one to four, whereas discussion opportunities are rated as present or not present. Finally, there are master rating standards. In order to move from behaviours in a classroom to words on a page to the numerical rating that can be aggregated and analysed, video examples of what counts in the scales and rating points are needed. Master raters, individuals who have a deep understanding of the observation scales and rating points, create the “gold standard” master ratings to which raters’ understandings of the scoring tools are compared. This is done by having master raters assign master ratings to many lessons and segments of lessons. By master rating videos to show, for example, what a two versus as three looks like for the routines component, master ratings show raters how to apply the rating scales. Master ratings establish the behavioural standards that raters use as benchmarks to guide their application of rating scale words to lesson behaviours. Such master coded videos were critical to developing a standardised interpretation of the Study’s codes.

Table 4.2. Observation system components

Rating specifications	Rating processes	Sampling and scoring specifications
Aspects of teaching Scales Standards in master ratings	Training Certification Calibration Validation Rater assignment	Student sampling Time sampling Subject matter sampling Scoring model

Source: Liu et al. (2019, p. 65[10])

Rating processes. To ensure that raters can reliably and accurately apply the rating specifications, all observations systems use rating processes – although the degree to which a single observation system uses all of the rating process components varies widely (Bell et al., 2019[9]).

Training materials were developed to help raters understand the aspects of teaching measured in the system as well as the rating scales. Many systems require raters to become certified on the system, a process that ensures raters have learnt how to apply the rating specifications during formal training activities. After raters begin rating lessons, regular calibration events may provide opportunities for raters to all rate the same video, get feedback on their ratings, and continue to learn and calibrate their application of the rating specifications. Validation videos may be used when rating is done by video. Validation videos are pre-rated by master raters and put into the rater’s rating queue. To the rater, validation videos look just like any other video that is a part of the Study so the rater is not aware they are being evaluated. However, validation videos are used to determine the rater’s degree of fidelity to the rating specifications when carrying out routine ratings.

Videos are frequently rated by more than one independent rater. Multiple ratings are used to calculate metrics of inter-rater agreement. And finally, rater assignment is the process used to assign raters to teachers, schools or classrooms so that no one rater has an undue impact on the ratings for that unit of analysis. For example, if four videos from a single teacher must be rated, raters would be assigned to those four videos to maximise the number of different raters that rate those four videos. Careful rater assignment may lead to higher quality scores.

Sampling and scoring specifications. Because almost no study will ever observe and rate the universe of teaching for a year, observation ratings are bounded by a sample of the teaching a group of students experience over some period of time. Observation systems must then employ sampling and scoring specifications. For example, for teachers who teach multiple groups of students, observation systems specify how groups are sampled. This may mean selecting a single focus classroom group from six groups of seventh-grade students, or it might mean selecting one group of students from each grade taught by the teacher.

Within any group of students, interactions with teachers happen over time – a lesson, a unit, a semester, a school year. Observation systems must delineate how ratings will sample time within and across lessons. Will the rater watch the whole lesson and then rate? Will they rate smaller intervals of a lesson? Are the lessons spread across an entire school year? A semester? In some cases, teachers teach different subjects. Whether a rater sees lessons in all of those subjects or in a single one, subject matter sampling is also specified by the observation system. And finally, observation systems must determine how ratings will be aggregated and weighed in scoring models. Scoring models shape how the observation scores operate in study analyses and should be considered carefully (Hill, Charalambous and Kraft, 2012[8]).

Two types of codes in the Study

In both the research literature on teaching quality and in existing observation systems, it was clear that teaching and learning behaviours needed to be tracked and coded in different ways depending on the time scale of the behaviours being rated (Bell et al., 2019[9]) the grain-size of the behaviours (Brophy et al., 1986[11]; Hill and Grossman, 2013[12]) and the level of judgement required from observers. Accordingly, two types of codes were defined to capture variation in teaching and learning behaviours:

- **Indicator codes.** These codes capture whether a particular behaviour happened or not (e.g. working in pairs or in a whole class structure, using a calculator) as well as the quality of specific small fine-grain- sized practices (e.g. the explicitness of the lesson's purpose or the strength of a mathematical summary).
- **Component codes.** Component codes capture the level of quality of a specific teaching construct for which behaviours occur over longer periods of time and are at more coarse grain-sizes (e.g. the quality of questioning in a lesson, or the quality of teachers' feedback to students).

Each code (indicator or component) has associated behavioural examples with a descriptor for each score point. It also has 1-4 video anchors. A video anchor is a brief video clip that shows examples of how the code descriptions map onto classroom behaviours. While the practices are generalizable, the specific examples provided for each construct were based on mathematics and more precisely on the focal unit. Tables Table 4.3 and Table 4.4 illustrate an indicator and a component code for the domain "Quality of subject matter".

Table 4.3. Example of an indicator from the 'Quality of subject matter' domain

Explicit learning goals	1 Little explicitness	2 Some explicitness	3 Predominantly explicit
The extent to which the teacher poses explicit learning goal(s) to students for the lesson and activities.	The teacher does not explicitly state or write the learning goal(s) or activities.	The teacher explicitly states or writes the activities or topic(s) in which students will engage. There is no explicit statement of the learning goal(s).	The teacher explicitly states or writes the learning goal(s).

Source: OECD, Global Teaching InSights Database.

Table 4.4. Example of a component from the 'Quality of subject matter' domain

Explicit patterns or generalisations	1	2	3	4
The teacher or students look for patterns in their work together. They also generalise from the specific work the students are working onto a foundational concept and/or definitions underlying the specific work.	Neither the teacher not students look for patterns in the mathematical work. OR They do not generalise from the work.	Teacher looks for patterns in the mathematical work. Identified patterns focus on surface features of the mathematics. OR Explicit generalisation(s) are developed from the mathematics under consideration and focus on nomenclature or algorithm processes. They are muddled, correct or incorrect, and superficial.	Students look for patterns in the mathematical work. Identified patterns focus on surface features of the mathematics. OR Explicit generalisation(s) are developed from the mathematics under consideration and focus on nomenclature or algorithm processes. They are clear, correct, and elaborated. If they generalise to foundational concepts, ideas, and/or definitions, the generalisations are somewhat muddled.	Teacher or students look for patterns in the mathematical work. Identified patterns focus on surface features of the mathematics. OR Explicit generalisation(s) are developed from the mathematics under consideration and focus on nomenclature or algorithm processes. They are clear and correct.

Source: OECD, Global Teaching InSights Database.

As described earlier and in Table 2.3 in Chapter 2, each component includes a holistic, or "overall", domain rating. Given the feasibility goal of the Study, raters also assigned a holistic domain rating using the information gathered for the component ratings. The holistic ratings will be compared to the aggregated component ratings to determine if there are more desirable psychometric properties of either the aggregated or holistic domain ratings. Raters received instructions about how to rate the holistic domain rating; these are provided in Annex A. Indicator codes and associated materials are shown in Annex B. The similarities and differences between these types of codes are further discussed in the next section.

Important aspects of the Study's observation coding system

The Study's rating specifications – the aspects of teaching that are measured by indicator and component codes and the scales used to measure them are explained thoroughly in Annex A (Components) and Annex B (Indicators).

Master ratings – the "gold standard" numerical ratings for specific videos – were finalised for all training, certification, calibration and validation videos from May 2018 to September 2018. Annex A and Annex B include the list of training and certification videos that were used in component and indicator training as well.

Videos that were master rated were acquired through the pilot study. During pilot data collection, 12 teachers in each country/economy had one lesson video-recorded. Transcripts and artefacts associated with each video were created in a two-stage process described in Chapter 11. From May to September, each video was rated by both indicator and component codes. Master raters (at least two raters from different country/economies per observation code) then reviewed and revised the initial master ratings. The observation team reviewed all suggested revisions and finalised the master ratings with the finalised set of codes. These finalised master ratings were used in the master rating training held in Pittsburgh, the

United States (October, 2018). The Study's coding and training materials were used to certify both master raters and country/economy raters.

References

Bell, C. et al. (2019), "Qualities of classroom observation systems", *School Effectiveness and School Improvement*, Vol. 30/1, pp. 3-29, http://dx.doi.org/10.1080/09243453.2018.1539014. [9]

Bell, C. et al. (2015), "Improving observational score quality: Challenges in observer thinking", in Kane, T., A. Kerr and K. Pianta (eds.), *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project*, John Wiley & Sons, San Francisco, CA., http://dx.doi.org/10.1002/9781119210856.ch3. [5]

Brophy, J. et al. (1986), "Teacher behavior and student achievement", in Wittrock, M. (ed.), *Handbook of Research on Teaching (3rd edition)*, Macmillan, New York, NY. [11]

Casabianca, J., J. Lockwood and D. McCaffrey (2015), "Trends in classroom observation scores", *Educational and Psychological Measurement*, Vol. 75/2, pp. 311-337, http://dx.doi.org/10.1177/0013164414539163. [6]

Floman, J. et al. (2017), "Emotional Bias in Classroom Observations: Within-Rater Positive Emotion Predicts Favorable Assessments of Classroom Quality", *Journal of Psychoeducational Assessment*, Vol. 35/3, pp. 291–301, https://doi.org/10.1177/0734282916629595. [7]

Hill, H., C. Charalambous and M. Kraft (2012), "When rater reliability is not enough", *Educational Researcher*, Vol. 41/2, pp. 56-64, http://dx.doi.org/10.3102/0013189X12437203. [8]

Hill, H. and P. Grossman (2013), "Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems", *Harvard Educational Review*, Vol. 83/2, pp. 371-384, http://dx.doi.org/10.17763/haer.83.2.d11511403715u376. [12]

Linn, R. (1994), *Assessment-Based Reform: Challenges to Educational Measurement*, Educational Testing Service, Princeton, NJ., https://eric.ed.gov/?id=ED393875 (accessed on 8 January 2020). [3]

Liu, S. et al. (2019), "Classroom observation systems in context: A case for the validation of observation systems", *Educational Assessment, Evaluation and Accountability*, Vol. 31/1, pp. 61-95, http://dx.doi.org/10.1007/s11092-018-09291-3. [10]

Mislevy, R., L. Steinberg and R. Almond (2003), "Focus article: On the structure of educational assessments", *Measurement: Interdisciplinary Research & Perspective*, Vol. 1/1, pp. 3-62, http://dx.doi.org/10.1207/S15366359MEA0101_02. [2]

Paine, L., S. Bloemeke and O. Aydarova (2016), "Teachers and teaching in the context of globalization", in Drew H. Gitomer and Courtney A. Bell (eds.), *Handbook of research on teaching*, American Educational Research Association, Washington DC. [1]

Shepard, L. (1989), "Why we need better assessments", *Educational Leadership*, Vol. 46/7, pp. 4-9. [4]