Global Teaching InSights

Technical Report

Section II: Instrument development

6 Rating teaching components and indicators of video observations

Courtney A. Bell

To ensure that video scoring scales meant the same thing within and across countries/economies, the Study established a coherent video rating process. The chapter describes the development of materials for rating video recordings of teaching using the components and indicators. It also explains the training process for master raters and raters in each participating country/economy.

Introduction

Global Teaching InSights (results from the TALIS Video Study project, and is hereafter cited in this chapter as "the Study" or "GTI") uses standardised procedures for training and certifying video raters, and for coding videos in every participating school system. This is important because in studies with less stringent processes it can be challenging to determine whether differences across countries are real or simply the result of variation in implementation.

For ensuring the coherence of video rating processes across participating countries and economies, an essential step was to develop English-language training materials, which were later used consistently for training raters and coding videos. These materials were the same materials used in the preparation of global master raters, who were responsible for training other raters in their respective country/economy. This chapter describes the video rating processes for the Study and the challenges to establishing well-defined training procedures across countries/economies.

Development of training materials

Four major principles guided the development of the training materials: 1) coherent structure to ensure the scoring scales meant the same thing within and across countries/economies; 2) equitable country/economy representation so that the raters developed a broad understanding of how teaching and learning looked globally; 3) robustness to educate bilingual raters who had a mathematics background in quadratic equations but were not mathematics experts; and 4) explicitness to be effective in a face-to-face training approach for all raters. The observation team would train the country/economy master. raters (hereafter master raters) face-to-face, and the master raters would train the raters face-to-face in their own countries/economies.

As mentioned in Chapter 5, all master-rated training materials were intended to be developed from lessons recorded in the 12 pilot teachers' classrooms in each country/economy. As Annex 6.A shows, 125 pilot videos were submitted. Many of these videos were unusable because they did not pass quality control checks for audio and video quality or they were not focussed on the topic of quadratic equations. In these cases, videos from the main study were collected and master rated to ensure equitable representation of country/economies in training, certification, calibration, and validation videos (see Annex 6.A).

The Study's observation code development team (hereafter the observation team) produced two separate sets of training materials: one for components and one for indicators. The distribution of country/economy videos across training videos is shown in the component and indicator video inventories (see Annex A.5 and B.5, respectively). Twenty-four videos were used in the component training materials and 23 were used in the indicator materials. These totals do not include the additional videos used for calibration and validation videos – another 56 videos. In all, every country/economy contributed between 12-13 videos to those that were master rated. Some videos were used for multiple purposes, resulting in a total of 99 videos that were master rated and used in training, certification, calibration, or validation activities. All master rated videos were subtitled in English.

The clips of the training videos selected by the observation team varied in length, depending on their purpose. Benchmark clips – clips selected to define a scale point on a specific code – ranged in length from 0:35 to 23:07 minutes, depending on the code. For the actual rating process, all videos were segmented to improve rater accuracy and decrease cognitive demand. Raters were allowed to use the artefacts collected from that day's lesson when it was difficult to understand what was in the video. For teaching components, videos were rated in 16-minute segments. Teaching indicators were rated in 8-minute segments. Segmenting rules for videos that did not divide evenly into 8- or 16-minute segments are included in Annexes A.2 and B.2. Training practice videos were usually one segment long (16 or 8 minutes, depending on the codes being rated). Calibration videos were always 16 minutes long and

therefore had one (components) or two (indicator) segments. Certification videos are described below. For a review of each of these video types used to support the rating processes of the observation system, see Chapter 5.

Indicator training materials

The indicator training materials (see Annex B) included the following:

- **Codes.** Rubrics for each indicator on the various scales and general rules for applying the codes. Raters were expected to consult these rubrics when rating.
- **Training manual.** A reference manual with rules, heuristics, guiding questions, more detailed definitions and written examples. Raters were expected to consult this manual when rating.
- **Training slides for raters.** This PowerPoint slide deck organises raters' learning into groups of codes. Slides explain video-based benchmarks for the varied rating scales as well as practice video clips used to provide raters formative feedback on their learning. Video clips are not embedded in the slides to adhere to human subjects protections. Raters were expected to refer to these slides and videos when rating.
- **Training slides for master raters.** This PowerPoint slide deck is the same as the raters' deck with the addition of content that teaches master raters how to manage training materials and the learning process of the raters. Slides were used by the observation teamwith master raters during the train-the-trainer sessions and were expected to be used as reference material when master raters trained raters.
- **Indicator video inventory.** This document shows which videos are benchmarks and practice videos. There were 48 total video clips included in the training. These were used to support training learning objectives and as a reference during rating.
- **Indicator training handout.** This handout is an annotated transcript of a mathematics discussion as an example of the "discussion" indicator code. It was used to support learning during training and as a reference.
- **In person training agenda.** This agenda was used to train master raters in an in-person meeting between 20 October and 29 October 2018, in Pittsburgh, the United States and describes each day of activity for indicator training. Master raters were allowed to break the agenda into shorter training days when training raters; however, any change in the order of training slides or videos had to be reported to the International Consortium
- Certification tests. There were two indicator certification tests for the master raters. The first was comprised of one lesson from each of two countries/economies. Lessons varied in length; the two lessons had a total of eight segments. The second test was structured in the same way with two lessons from a third and fourth country/economy and a total of six segments.

Component training materials

Much like the indicator training materials, component training materials (see Annex A) included the following:

- **Codes.** Rubrics for each component on the 4-point scoring scale and general rules for applying the codes. Raters were expected to consult these rubrics when rating.
- **Training manual.** A reference manual with rules, heuristics, guiding questions, more detailed definitions and written examples. Raters were expected to consult this manual when rating.
- **Training slides for raters.** This PowerPoint slide deck organises raters' learning into groups of codes. Slides explain video-based benchmarks for the 1-4 rating scales as well as practice clips used to provide raters formative feedback on their learning. Video clips are not embedded in the

slides to adhere to human subjects protections. Raters were expected to refer to these slides and videos when rating.

- **Training slides for master raters.** This PowerPoint slide deck is the same as the raters' deck with the addition of content that teaches master raters how to manage training materials and the learning process of the raters. Slides were used by the observation team with master raters during the train-the-trainer sessions and were expected to be used as reference material when master raters trained raters.
- **Component video inventory.** This document shows which videos are benchmarks and practice videos. There were 52 total video clips as a part of the training. These were used to support training learning objectives and as a reference during rating.
- **Component training handout.** This handout elaborates the definition and examples of what is meant by a "connection" for the "explicit connections" component code. It was used to support learning during training and as a reference.
- **In person training meeting agenda.**
- This agenda was used to train master raters in an in-person meeting and describes each day of activity for component training. Master raters were allowed to break the agenda into shorter training days when training raters; however, any change in the order of training slides or videos had to be reported to the International Consortium.
- **Certification tests.** There were two certification tests for all raters. The first was comprised of one lesson from each of two countries/economies. Lessons varied in length; the two lessons had a total of five segments. The second test was structured in the same way with two lessons from a third and fourth country/economy and a total of four segments.

Training master raters

Because many of the country/economy experts assumed master rater roles, it is best to consider master rater training as comprised of two parts: 1) the development work in which the master raters participated (as described above); and 2) formal training. Through the four iterations of development work an initial shared understanding of the observation constructs and codes was built. All countries/economies were required to have at least two master raters; Germany*[1] and Mexico each had four. There could be a single master rater team that served for both components and indicators or two master rater teams that oversaw components and indicators separately.

The formal training of master raters was conducted by members of the observation team. The in-person sessions took place in Pittsburgh, the United States from 22-29 October 2018. Indicator code training and certification occurred first followed by component code training and certification in the latter part of the week. The detailed schedules are provided in Annex A.7 and Annex B.7.

Prior to indicator and component training sessions, master raters attended a voluntary half-day meeting that discussed the common issues and considerations all country/economies needed to address as they carried out their own rater training and main study rating. This included the specification of rater processes such as certification, calibration, validation, multiple ratings and rater assignment.

Master rater indicator training

Indicator code training comprised two training days and one certification day. It began with a bias activity. During this activity, master raters were asked to reflect on specific times in watching classroom videos where they were likely to feel their bias appear. For example, a teacher with a flat affect tends to make some raters perceive a lack of social-emotional support in that classroom. Master raters were asked to note their biases and opinions as well as their personal views of good teaching and, to the best of their

ability, set them aside to focus on their learning goal for the remainder of the training. That goal was to learn how to apply the Study's conceptualisation of good teaching and to learn how to teach their own raters to use the indicator codes.

Training activities were guided by four research principles of how people learn activities were designed to be learner-centred, knowledge-centred, assessment-centred, and community-centred (National Research Council, 2000, pp. 3-30[1]). Table 6.1 shows how these global research principles map onto the specific activities in the training materials detailed in Annexes A and B.

Table 6.1. Learning principles used for the Study's teaching indicator and component rater training

Learning principle	Definition of principle	Specific approaches in indicator and component training materials
Learner-centred	Pay careful attention to the knowledge, skills and attitudes the raters bring into the classroom	Clear delineation of learning goals and expectations Bias training
Knowledge-centred	Focus on what is taught, why it is taught and what mastery looks like	Broke rating into sub competencies: using only evidence to make decisions, knowing and using the TVS code and scale language, and reasoning with the rating scales Reasoning strategies included rules and definitions, as well as metacognitive strategies (guiding questions and heuristics) that were general and code-specific Many video examples of codes that provided definitions and master ratings
Assessment-centred	Ongoing formative assessments designed to make raters' thinking visible and adjust training accordingly	Raters find evidence and compare it to master-rated evidence regularly Raters have early and frequent practice opportunities Optional additional practice opportunities (if needed by raters)
Community-centred	Development of classroom norms	Relentless classroom focus on evidence (what was said or done in the video), TVS code language use and rater reasoning All raters sharing their rating efforts with one another and the group

Source: OECD, Global Teaching InSights Database.

Each training day included similar activities: reading training manual notes, PowerPoint slides and handouts; listening to the facilitator present slides; viewing benchmark videos on Kaltura (a video-sharing platform); taking notes that identify evidence to use in coding (or annotating the video transcripts while watching videos); whole group and smaller group discussion of codes and videos; and formative practice with the codes on both segment-length and shorter video clips.

The following strategies were used to help master raters. The training manual provided guiding questions to focus raters' attention on certain behaviours or evidence, features that should play an important role in the rater's thinking process when deciding on a score, rules and heuristics, as well as narrow guidance about where and how to count specific evidence. The training also provided explicit cognitive guidance for and examples of what to do when common rating problems arose such as evidence splitting over

descriptors and how to reason a final score or not having enough evidence for a specific indicator. Explicit examples of how not to reason or apply the codes were also an important part of the training, which allowed master raters to practice reasoning correctly under deliberately challenging circumstances. Finally, benchmark-training videos were used to show how different behaviours might receive the same rating on an indicator with different video evidence to support that rating.

At the end of the two days, master raters took a certification test for which they scored two full-length quadratic equation lessons. To pass the certification test, master raters had reach a certification standard. They had to agree exactly with at least 75% of the indicator master ratings and agree exactly or adjacently[2] for 90% of the master ratings. All master raters achieved the established standard of accuracy on the first try and did not need to receive additional support from the observation team, nor take a second certification test. Exact and adjacent agreement varied somewhat across country/economy master raters; however, across all master raters the average exact agreement was 86% and adjacent agreement was 97% (Table 6.2).

Table 6.2 . Master rater indicator certification agreement rates, by country/economy

Country/economy	Number of master raters tested	Percentage that passed on first attempt	Percentage of overall average exact agreement	Percentage of overall average exact and adjacent agreement
B-M-V (Chile)	2	100	90.0	97.8
Colombia	2	100	86.0	97.3
England (UK)	2	100	88.5	97.5
Germany*	4	100	84.5	96.9
K-S-T (Japan)	2	100	87.0	95.8
Madrid (Spain)	2	100	88.5	97.3
Mexico	4	100	81.3	96.0
Shanghai (China)	2	100	84.0	98.3
Total/average	20	100	85.6	96.9

Notes: Bíobio, Metropolitana and Valparaíso (Chile) (hereafter "B-M-V [Chile]").
Kumagaya, Shizuoka and Toda (Japan) (hereafter "K S T [Japan]").
*Germany refers to a convenience sample of volunteer schools.
Source: OECD, Global Teaching InSights Database.

Master rater component training

Component code training lasted three training days plus two additional hours. Certification was one day and there was one "break day" between the end of training and the certification test. Some master raters used this day for preparing for the certification test.

The activities used in the component training were the same as for indicators (refer to the description in the previous section for details) with one exception: For master rater training, the observation team did not repeat the bias activity; however, when master raters returned to their own country/economies, they did do the bias activity to begin the component training.

The same rater teaching and learning approach was taken with components as previously described with indicators (Table 6.1).

Just as with indicators, master raters had to pass a components certification test after the three days of training. It too comprised two full-length quadratic equation lessons. To pass the certification test, master raters had to reach specific certification standards. Specifically, they had to agree exactly with at least 50% of the component ratings and agree exactly or adjacently with at least 85% of ratings.

All master raters achieved the established standard of accuracy on the first try and did not need to receive additional support from the observation team or take a second certification test. Country/economy results are listed in Table 6.3. Across all country/economy master raters, the average exact agreement was 62% and exact plus adjacent agreement was 97%.

Table 6.3. Master rater component certification agreement rates, by country/economy

Country/economy	Number of master raters tested	Percentage that passed on first attempt	Percentage of overall average exact agreement	Percentage of overall average exact and adjacent agreement
B-M-V (Chile)	2	100	60.0	96.5
Colombia	2	100	67.5	97.0
England (UK)	2	100	70.5	97.5
Germany*	3	100	60.0	95.0
K-S-T (Japan)	3	100	58.0	96.3
Madrid (Spain)	2	100	64.0	98.0
Mexico	4	100	60.0	96.5
Shanghai (China)	2	100	60.0	96.5
Total/average	20	100	61.9	96.6

Note: *Germany refers to a convenience sample of volunteer schools.
Source: OECD, Global Teaching InSights Database.

Training raters in participating countries/economies

Master raters were given access to all training videos through Kaltura and provided with electronic copies of all revised English language training materials. Revisions reflected small changes identified during the master rater training including the correction of typographical errors, the addition of one benchmark video agreed upon during master rater training and handouts.

Master raters were expected to carry out indicator and component training separately with their county/economy raters, exactly as training was conducted for them (e.g. face-to-face, in the order specified by the agenda, using the slides provided). Minor changes to the training materials had to be documented and approved by the International Consortium (Table 6.4). Some countries/economies translated specific training materials (e.g. the codes) into their country/economy languages. This was permitted; however, the IC warned master raters to be extremely careful not to alter the very specific meanings associated with the English version of the training materials. Any errors or discrepancies in translation might have led to misunderstandings of the codes which could, in turn, have led to lower performance on the quality control rating processes of certification, calibration, validation and double-rating. Country/economy raters were still expected to use the English language materials during training and as their main reference materials when rating even if a translated version was available.

Table 6.4. Indicator or component training modifications, by country/economy

Country/economy	Work Prior to Training	Translation	Training
B-M-V (Chile)			
Colombia			
England (UK)			Training schedule had to be broken up into shorter training days to accommodate rater schedules.
Germany*	All raters attended a one day knowledge workshop on quadratic equations taught by the national team's mathematics expert.		There were two components training sessions after the original raters were certified in order to have enough raters to complete ratings on the study timeline.
K-S-T (Japan)		Translated all training materials and codes.	One additional training video was used with raters for extra practice.
Madrid (Spain)	Orientation session to the overall project and video watching session.		
Mexico			
Shanghai (China)		Translated all training materials and codes.	

Note: *Germany refers to a convenience sample of volunteer schools.
Source: OECD, Global Teaching InSights Database.

The implementation of country/economy indicators and component training is described in Chapter 18. The country/economy main study rater certification process results, calibration and validation results are reported in Chapter 18.

References

National Research Council (2000), *How People Learn: Brain, Mind, Experience, and School*, National Academies Press, Washington, DC., http://dx.doi.org/10.17226/9853. [1]

Annex 6.A.

Annex Table 6.A.1. Country/economy video submissions and master rater coding: Pilot and main study

Country/economy	Pilot videos submitted	Pilot videos master coded	Main study videos master coded
B-M-V (Chile)	15	10	2
Colombia	14	8	4
England (UK)	12	10	3
Germany*	12	6	7
K-S-T (Japan)	14	11	2
Madrid (Spain)	19	5	7
Mexico	9	5	7
Shanghai (China)	30	12	0
Total	**125**	**67**	**32**

Note: *Germany refers to a convenience sample of volunteer schools.
Source: OECD, Global Teaching InSights Database.

Notes

[1] Germany* refers to a convenience sample of volunteer schools.

[2] There are different rating types. “Master ratings” refers to the "correct” ratings for a video. "Indicator/component master ratings” describes the rating allocated to the type of observation code. “Exact ratings” occurs when the rater and master rater assign the same rating to the segment. “Adjacent ratings” occur when the rater assigns a numerical rating 1 off from the rating the master rater assigned.