Organisation for Economic Co-operation and Development

EDU/EDPC(2019)9

For Official Use **English - Or. English**

22 March 2019

DIRECTORATE FOR EDUCATION AND SKILLS
EDUCATION POLICY COMMITTEE

THE OECD TALIS VIDEO STUDY – Progress Report

11-12 April 2019

This document provides an update on the progress made so far in the design and implementation of the project. Delegates are invited to NOTE the progress report on the TALIS Video Study.

Karine Tremblay, Senior Analyst (Tel.: +33 1 45 24 91 82, email: karine.tremblay@oecd.org), Ms Anna Pons, Analyst, (Tel.: +33 1 45 24 91 87, email: anna.pons@oecd.org) Directorate for Education and Skills

JT03445173

Video Study Progress Report

Background

1. The TALIS Video Study of Teaching Practices was initiated in the 2015-16 PWB and is conducted under the auspices of the Education Policy Committee.

2. As the data collection phase has now reached completion and the study is entering its final analysis stage, it is timely to update the Education Policy Committee on the progress made so far in the design and implementation of the project, and present its next steps. The Education Policy Committee is invited to:

- **NOTE** the progress report on the TALIS Video Study.

The rationale behind the TALIS Video Study

3. The overarching goal and rationale of the TALIS Video Study is to trial new methodologies to deepen our understanding of teaching and learning at an international scale. To have a more rounded picture of the classroom, the Study collects observation and artefact evidence in addition to survey and achievement data.

4. Most research on teaching effectiveness in recent decades has been based on indirect classroom data (Grossman et al., 2014; Hill, Kapitula and Umland, 2011; Pianta and Hamre, 2009). These studies have provided important insights on potential factors influencing teaching and learning practices (Wayne and Young, 2003). These research methods are incapable of providing, however, the detailed information about a teacher's organisational and pedagogical processes that is actually needed to inform and improve teaching practice in the classroom (Grossman et al., 2014; Pianta and Hamre, 2009).

5. Without a richer understanding of classroom practice, what constitutes effective teaching can only be defined narrowly. The prevailing view is that a "good teacher" is one able to produce student gains, and a 'bad teacher' as the opposite (Pianta and Hamre, 2009). Of course, defining an effective teacher only by results and not by the quality of practice does not contribute to the development and continuous improvement of educators.

6. OECD work, notably PISA and TALIS, has provided valuable comparative insights about teaching and learning across the world. In PISA, students are asked about the practices used by their teachers. In TALIS, teachers are asked about their teaching practices as well as their working conditions, attitudes, and beliefs. Furthermore, recent efforts to link the TALIS and PISA studies have allowed for further exploration of teaching practices at the level of schools.

7. The TALIS Video Study aims to complement already established OECD studies in a twofold way:

- First, the Study will provide more objective evidence on classroom processes by drawing on direct measures of classroom teaching and instruction. Using teachers'

self-reports to measure instructional quality is particularly challenging because these reports frequently reflect responses that the teachers consider socially desirable (Little, Goe and Bell, 2009; van de Vijver and He, 2014). By looking directly into the classroom through observation and artefact evidence, the TALIS Video Study can overcome the limitations of self-reported data and validate existing measures of practice.

- Second, the TALIS Video Study will enhance our understanding of what happens in classrooms around the world in a comparative manner. with richer information about classroom processes. Though more challenging from a methodological standpoint, observational studies can provide a deeper and richer understanding of pedagogical processes and practices in classrooms, and improve our understanding of how these are related to student learning and other outcomes. Observation is perhaps the most powerful way of measuring the so-called tacit or sticky knowledge that underpins quality classroom practice, even if it can be difficult to formalise and isolate this knowledge from the place where it is created and applied.

8. Artefacts (e.g. lesson plans, in-class assignments, and student homework) can provide a window into classroom practice without being as intrusive and labour-intensive as observation. There has been promising recent research on the possibilities of making reliable judgments about instructional quality in a number of subjects based on the classroom artefacts (Borko, Stecher and Kufner, 2007; Matsumura et al, 2002; Matsumura et al., 2006; Martinez et al., 2012; Stein and Lane, 1996).

The goals of the TALIS Video Study

9. The TALIS Video Study (TVS) aims to trial new methodologies to investigate teaching practices to support a deeper understanding of teaching at both national and global levels. Concretely, the study is designed to:

- understand which aspects of teaching are related to student learning and student non-cognitive outcomes
- observe and document how the teachers from participating school systems in the study teach
- explore how various teaching practices are inter-related, and how contextual aspects of teaching are related to the student and teacher characteristics.

10. By repeating these types of analyses across participating school systems, the TVS will be able to identify both common and differing patterns in the results across the participating school systems. These findings will stimulate an increasingly nuanced discussion around teaching practices and student outcomes among educators, policymakers, researchers, and the general public.

11. This Study is also an important example of rich, innovative research contributing to the global understanding on education. By designing and applying new ways of measuring teaching and learning, the TALIS Video Study will make a valuable contribution to research across multiple contexts at scale. It will investigate the feasibility of various procedures for capturing teaching practices, such as video and artefact-based observations, and collecting data about the validity and reliability of alternative measures of teaching. In addition, the TALIS Video Study will provide important perspectives on the validity of self-report data because it will collect both survey and video data, offering a rich opportunity for triangulation.

The outputs of the Study

12. The expected outputs of the TALIS Video Study are:

- *Policy report:* describing the main findings of the study through tables, figures, and explanations of the relationships found. These insights will be useful for within-country stakeholders as they make policies and recommendations about diverse aspects of education and quality teaching.
- *Technical Report:* documenting all of the technical aspects of instrument and protocol development and of the methods followed in the analysis. It will also present the lessons learnt whilst navigating methodological, legal, technological, and implementation challenges to realise the ambitious goals of the Study.
- *Dataset:* a rich dataset of teaching measures and student achievement from its students and teachers to enable further explorations of teaching and learning.
- *New instruments to measure teaching:* the observation and artefact codes are internationally developed and validated rubrics for measuring teaching.
- *The Global Video Library:* a digital platform for creating and sharing expertise about teaching, including video-enhanced examples of the observation rubric and video clips that showcase the teaching practices that the Video Study finds to be most associated with student outcomes.

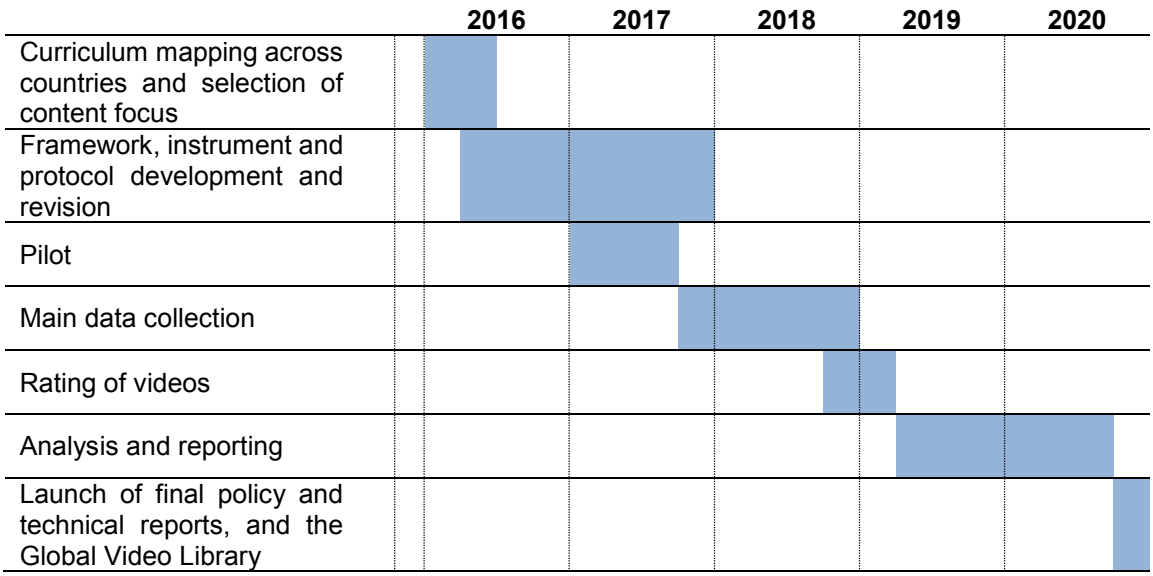
A major international effort

13. The TALIS Video Study is carried out on an international scale. It covers eight school systems, which feature a rich variety of classroom settings, pedagogical traditions, system-level policies, and student achievement levels. These systems are Chile, China (Shanghai), Colombia, Germany (8 *Länder*), Japan, Mexico, Spain (Madrid) and the United Kingdom (England). The United States also participated in the initial phases of the study.

14. This innovative OECD study is a major international effort. It involves national experts from different fields (i.e. pedagogy, survey methods and video observation) in each participating school system. Moreover, the Secretariat has contracted an international consortium of research organisations to implement the TALIS Video Study, led by RAND Corporation, and including Educational Testing Services (ETS) and the German Institute for International Educational Research (DIPF). Additionally, the Secretariat also set up a technical advisory group including 16 leading international experts in the measurement of teaching.

15. The implementation of the TALIS Video Study spans four years (Figure 1). The outputs of the Study will be launched by the end of 2020.

Figure 1. Timeline of the TALIS Video Study



16. At the time of writing in March 2019, the main focus of the work on this project follows:

- The rating of over 1000 hours of collected footage and its associated artefacts has been completed in all but one participating school system.
- After finalising the main data collection in October 2018 and testing the procedures for data submission, most participating school systems have successfully submitted all the data collected. The data verification and cleaning of country-level datasets is on-going.
- The preparations for the analytical and reporting phase have were also started. The Secretariat, the National Project Managers, the Consortium, and the Video Study Technical Advisory Group have reached agreement on how the data should be analysed and reported in the final report.

The design of the study

Key methodological aspects

17. The TALIS Video Study is trialling new education research methods at scale. The overall design builds upon existing research, and notably the TIMSS 1999 Video Study, the first and last video-based study at a large international scale (Box 1). However, significantly, the TALIS Video Study focuses on a common topic across countries, and uses standardized instrumentation and procedures for each school system. It looks at a diverse set of longitudinal measures of teaching and learning. The main design features are:

- *Common Evaluation Method*: Unlike many studies of teaching and learning, the TVS will draw on multiple measures of teaching at the same time to provide a more rounded picture of practice. The study develops a new set of instruments for evaluating video-recorded teaching practices and classroom artefacts, and fields them consistently in all participating countries.

- *Common Topic for Evaluation:* The Study focuses on the teaching and learning of a single common secondary mathematics topic (quadratic equations) to enhance the comparability across countries and the potential to capture the relationship between teaching and student outcomes. Using mathematics helps to reduce potential differences between countries in terms of curriculum or culture. Furthermore, using just a single topic also means that the focus is on how to teach, rather than what is being taught.
- *Longitudinal Design:* The Study captures student outcome measures before and after they have learnt the focal content, in order to take into consideration students' prior knowledge. Similarly, teachers and students are also surveyed twice to allow for full consideration of their contexts and perceptions.
- *Standardised Procedures*: The Study uses standardized and replicated procedures for data collection, for training and certifying video and artefact raters, and for coding videos and artefacts 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.

18. In spite of the standardisation of procedures, the Study is fully implemented by participating school systems. They bear the responsibility of fielding the data collection and rating aspects of the Study; for example, they are able to select their own recording and rater profile, as long as this produces data that meets common quality control standards. This local implementation will shed light on the feasibility, cost and scalability for implementing the Study in a larger number of countries in the future.

19. As with any truly innovative study, the TALIS Video Study is a proof-of-concept of new research methods to better understand teaching and learning.

20. It is worthwhile highlighting that the Video Study is not a global assessment of teachers or a ranking of countries' teaching quality. It will not provide a country mean or rank. Nor is it a comprehensive study of "the state of teaching", which would need far more school- and system-level contextual data, among other things.

21.

Box 1. Major differences between the TVS and the TIMSS Video Study

The TALIS Video study draws upon lessons learnt in previous observational international and national studies, in particular about instrumentation and data collection processes. Most notably, the "TIMSS Video Study". This attempted to draw on representative national samples of teaching "to provide a rich source of information regarding what goes on inside eighth-grade mathematics classes in Germany, Japan, and the United States". It was linked to the Third International Mathematics and Science Study (TIMSS) in 1995 and repeated in 1999.

The TIMSS Video Study produced some of the most valuable scholarship on critical issues in teaching and cross-cultural studies. It demonstrated the power of video to reveal classroom practices more clearly and subsequently facilitated the study of these. It also showed the potential video has to stimulate discussion and to deepen educators' understanding of teaching.

Nevertheless, although the TIMSS 1999 Video Study illustrated the promise of video, there were several important limitations that the TALIS Video Study has sought to overcome by:

- Measuring student learning gains: the foremost difference is that the TIMSS Video Study did not capture student learning following a longitudinal design. With no student learning gains data, it was not possible to connect features of teaching to student learning.
- Focusing on a curricula topic: The TIMSS Video Study did not call for any specific content to be covered; teachers were free to follow any goal and to teach any content that happened to be in the local curriculum. This made it particularly difficult to clearly comprehend differences in teaching methods as the content being taught often varied greatly across countries.
- Collecting two lessons per teacher: The decision to collect just one lesson per teacher in TIMSS prevented the study from considering some important aspects of teaching, such as the planning and reflection that takes place outside of the classroom.
- Developing observation protocols iteratively: The researchers in TIMSS let the heterogeneous teaching they observed—teacher talk, class assignments, teacher-student interactions, classroom movement, etc.—speak for itself, i.e., they developed their codes inductively, drawing both on prior research and the contents of the videos and artefacts they collected.
- Analysis of the quality of artefacts: In the TIMSS Video Study, lesson materials, such as classroom assignments, were used primarily to clarify the actions observed in the classroom videos; they were not scored independently.
- Building capacity for local implementation: The TIMSS Video Study procedures were established centrally and tightly controlled. For example, all raters were sent to a central location to score videos instead of building capacity for rating at the local level.
- Sampling a diversity of teachers: The use of national probability samples in TIMSS made it representative but also a particularly costly study.

Data collected

22. To obtain as complete a picture of teaching and learning as possible, the project collects a wide range of information from 85 teachers in each participating school system. All data related to that teacher and his or her teaching will be collected from one class/section of students. The following are the types of data that will be collected:

- *Video-Recorded Lessons*: Two lessons focused on quadratic equations will be video-recorded. To the greatest extent possible, these will be sampled so they are representative of the teaching of quadratic equations in that country. Some classroom practices depend on the function of the lesson, such as whether it is an introductory lesson or a practice lesson. Accordingly, for each teacher, the two lessons are randomly selected from one lesson in the first half of the unit and the second from the second half of the unit.
- *Instructional Artefacts*: Artefacts are collected from each lesson that is video-recorded and the lesson that follows it. These include lesson plans, instructional materials used during the lesson, homework assignments, or a copy of the next formal examination that includes quadratic equations so as to be able to understand the formal expectations for student understanding.
- *Student Pre- and Post-Test*: Students take a pre-test on their general mathematics knowledge two weeks before the start of the focal unit. They then take a post-test within two weeks of concluding the unit. The post-test is narrower in focus than the pre-test, in order to provide more precise measures of the students' knowledge and understanding of quadratic equations.
- *Student Pre- and Post-Questionnaire:* Students are asked about a range of aspects that can influence their learning before and after the focal units. This includes aspects such as family background, learning time both within and out of school, their perception of and participation in classroom activities, and their self-efficacy beliefs related to mathematics).
- *Teacher Pre- and Post-Questionnaire:* Teachers are asked about their background and education, their beliefs, their motivation, and their perception of the school environment. They are also asked about the selected class, the selected unit, including lesson goals, the mathematical content covered, the teaching practices used, and teachers' judgment of the effectiveness of the unit. Teachers are also questioned on whether the video-recorded lessons are representative of typical instruction. Moreover, teachers are asked to fill a "log" reporting on the number, length and activities for the lessons in which they teach quadratic equations throughout the focal unit.

23. In most participating countries, the focal topic is taught as part of the ISCED level 2 programmes; in Chile, the focal topic is taught in grade 11 at ISCED level 3.

Box 2. Piloting the instruments and data collection procedures

The instruments as well as the video and artefact fielding procedures were tested in a pilot that ran from January to July of 2017. The goal was to involve at least 12 teachers and 100 students in each participating school system to:

- assess the difficulty and item-functioning of the student test-items and questionnaire
- examine the time required to complete sections of the tests and questionnaires
- test the data collection of the teacher log
- test the capturing and processing procedures for videos and artefacts
- collect videos and artefacts that can be used to refine codes and training materials
- gauge the difficulty of recruiting participants for the study

Decision on focal unit based on curriculum mapping

24. The Study is using a single topic to study teaching and learning across multiple classrooms. A focus on learning within a single subject matter unit means that the Study's measures of teaching and learning (test, questionnaires, observation protocols) are tightly aligned. This will potentially provide greater power for testing the relationships between teaching practices and student learning.

25. Limiting the Study to a single curricular topic is important since studies such as the TIMSS Video Study (Stigler and Hibert, 2000) showed that without a consistent subject matter investigating both teaching and learning at the same time can lead to a conflation of both, and therefore does not yield clear patterns. This type of design has also been used successfully in previous research studies (e.g., Lipowsky et al. 2009).

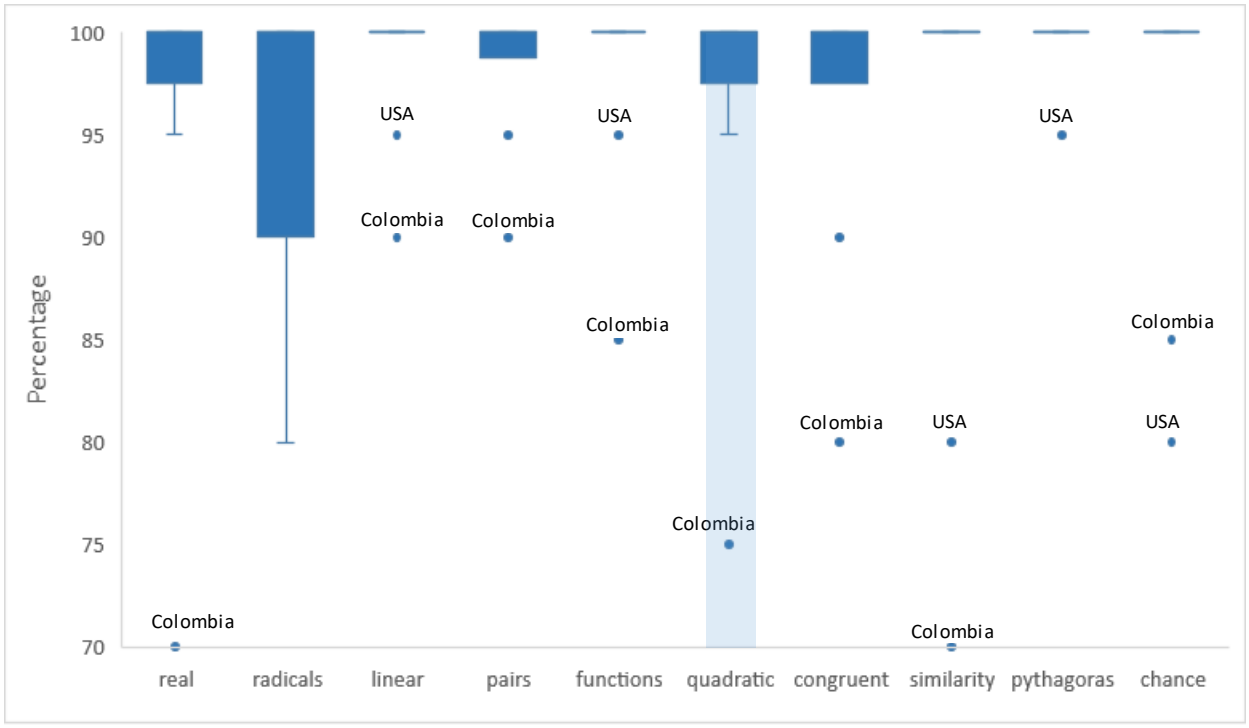
26. Furthermore, the topical focus on quadratic equations allows for more instructionally sensitive measures of teaching and student outcomes in a multi-country setting. Many of the measures that are used reflect a general vision of teaching quality and are widely generalisable.

27. The challenge in a global context is finding a common curricular topic. A curriculum mapping was thus undertaken at the outset of the Study to identify potential focal topics on the basis of the protocols of international assessments, previous comparative work on mathematics curricula, the curriculum for mathematics in the participating countries and economies, and textbooks used in participating school systems. The criteria for the selection of a focal topic for the Study included:

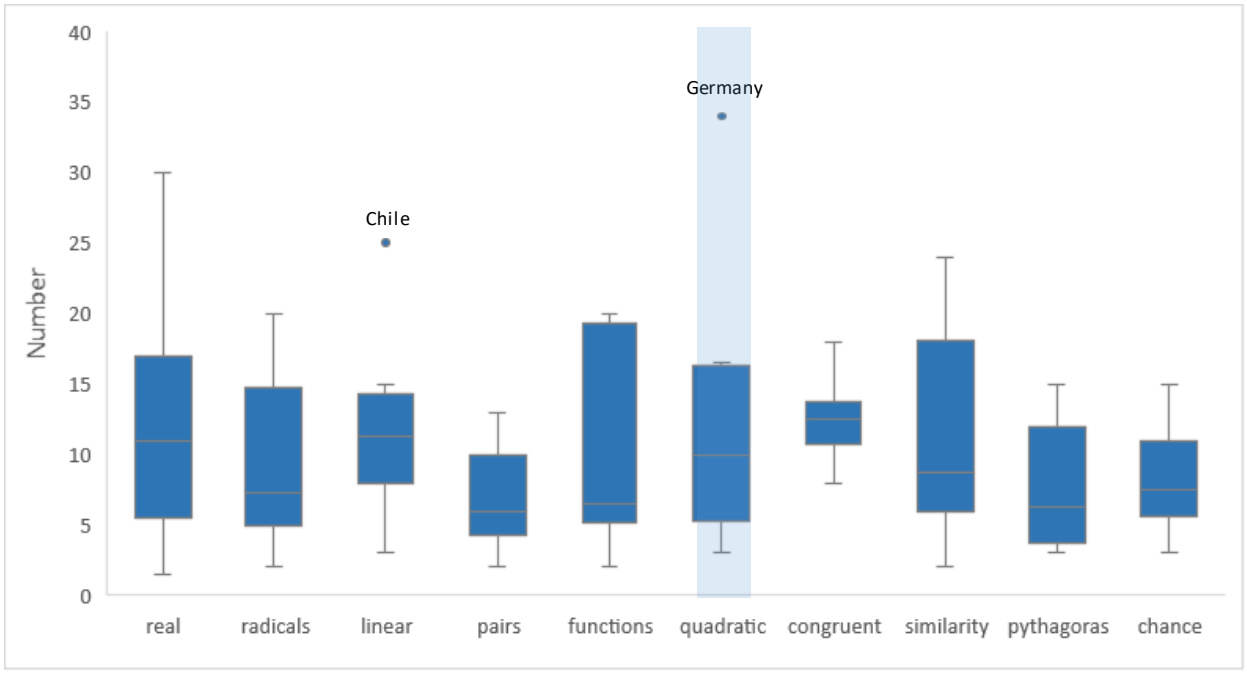
- Inclusion in the curriculum in all countries for essentially 100% of students
- Implementation close to age 15, in grade 8 or higher
- Consists of a reasonable number of lessons (6-12)
- Required a pedagogical approach with a well-defined starting point that introduces a core mathematical concept, whilst still having opportunities for rich mathematical activities (e.g., application, modelling and transfer), for deep mathematical thinking (e.g., argumentation, proof), and for working with different representations.

Figure 2. Variation across countries of selected curricular topics

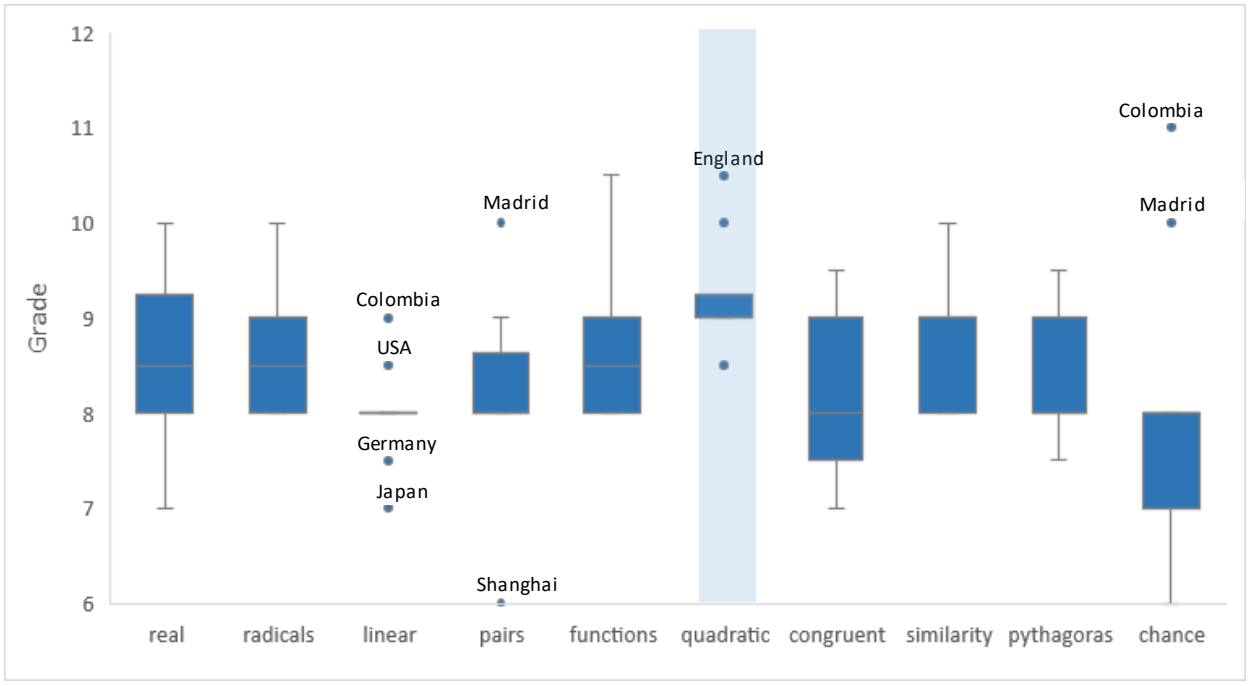
Percentage of students who are taught the topic



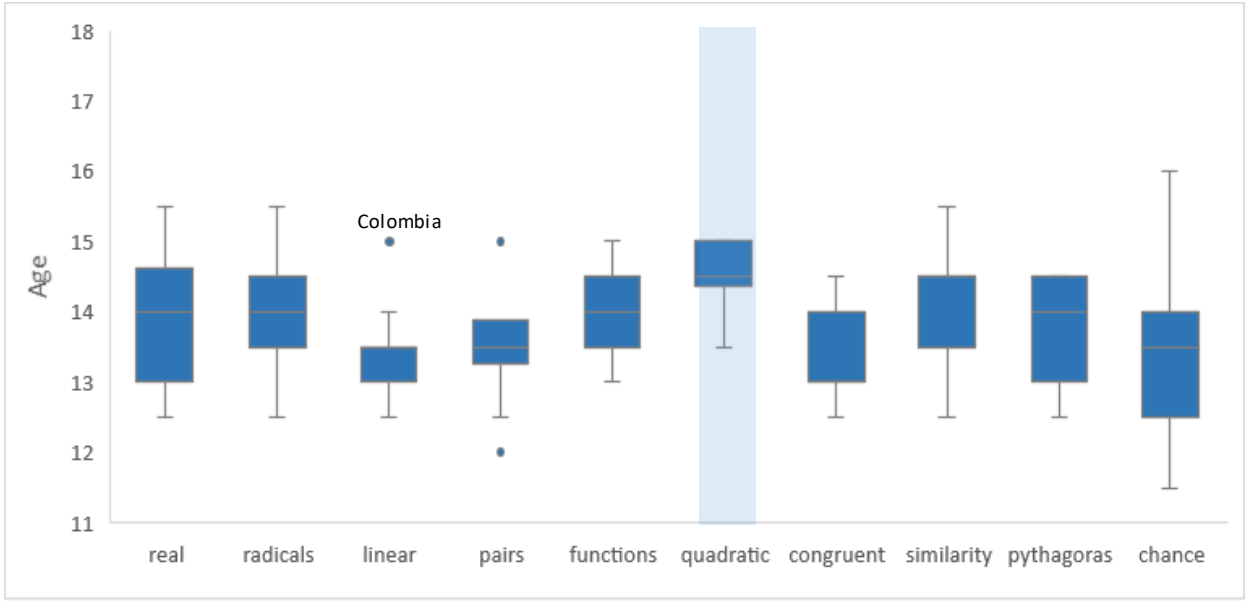
Number of lessons per topic



Grade level per topic



Student age per topic



28. Quadratic equations was chosen as the focal topic for the Study. The decision was made after considering that the age of students being taught quadratic equations would be closer to the PISA-target age of 15, there was less variation between countries on this topic compared to the teaching of simultaneous linear equations, and a common list of subtopics could be identified for quadratic equations. Also, students in PISA 2012 reported high levels of exposure to algebraic tasks, and experience with algebra, in particular with quadratic equations, was rather strongly correlated with achievement.

29. The curriculum mapping also revealed significant cross-country differences. For example, Asian countries (Japan, China) tend to teach algebraic operations independent from graphical material, while the United States, England, and Germany integrate algebra with graphical representations and everyday application. All countries implement a spiral curriculum across grades, but some split the curriculum up into small pieces even within a school year.

30. The mapping also revealed important differences in the implementation of the curricula as well. Most countries do not provide guidance for lesson planning, and the intended and implemented curriculum differ considerably. In some school systems, schools can teach the focal unit at any time throughout the year.

31. The differences observed confirm the findings of curriculum research carried out in the framework of previous international studies. These differences also provide a strong rationale for the more exhaustive curricular mapping being undertaken as part of the OECD Education 2030 project.

Sampling design and recruitment

32. The targeted sample of the TVS is 85 teachers who teach the focal unit from each country. This was calculated as the minimum for being able to observe enough variation in teaching and detect significant relationships with student outcomes.

A diverse sample of teachers

33. The TVS uses a stratified two-stage probability sampling design. This means that a teacher was selected to participate from the list of in-scope teachers (who teach the focal topic) for each of the randomly selected schools. One class and its associated teacher was randomly selected from the list. In addition, up to two replacement classes taught by other teachers were also selected. If the originally sampled teacher declined to participate, then a replacement teacher was invited to participate. If the replacement teacher declined, the second replacement teacher was invited. If this teacher refused, the sampled school was replaced by a replacement school.

34. As in other international surveys, participating school systems could choose to restrict the coverage of their national implementation of TVS to a subset of the geographic regions of the country. Attempting to survey teachers in schools located in geographically remote areas can be a costly, time-consuming, and statistically inefficient exercise.

The challenge of recruiting teachers and students

35. The reality is that the process has been mixed. Some school systems have managed to follow the random sample drawn (Shanghai, Mexico, Colombia, Madrid), even if exhausting most replacement schools (Chile). However, other school systems have deviated considerably from the random sample drawn. Japan and England have followed the random sample to some extent, but also recruited a small share of voluntary teachers and two teachers per school. Germany has not achieved the targeted number of 85 teachers, despite fully discretionally recruiting voluntary teachers and more than one per school (Table 1).

36. Participating school systems, in general, found that the recruitment of schools, teachers, and students took longer and was more challenging than expected. They exchanged good practices in the recruitment of participants in terms of planning, communication, incentives, and other efforts (Box 3).

Table 1. Number of participating schools and teachers

	Chile	England	Germany	Colombia	Mexico	Madrid	Shanghai	Japan
Schools	84	78	39	84	103	85	85	74
Teachers	84	85	50	84	103	85	85	89

Box 3. Good practices in the recruitment of participants

The sample size of the TALIS Video Study is considerably smaller than that of other OECD studies, yet the commitment required from each participant is larger. It entails entering their classrooms to observe them in action, asking them to fill in questionnaires and tests, and collecting their work through recurring visits to the school over several weeks. To increase the likelihood of success, participating school systems exchanged the following good practices:

- *Start recruitment early and keep track of participation:* A process of recruitment that began early with a robust, on-time system to track its progress, was highly recommended. England (United Kingdom), for example, developed an online system to track interest, consent levels, and engagement dates; appointed a co-ordinator to serve as the "face" of the study to build a strong and trustful relationship with schools; and prepared a FAQ list and counter-arguments for potential refusals.
- *Communicate the value and demands of the Study:* Brief and easy-to-read communication materials were developed to enable participants to understand the Study's value and what was expected of them. This included information booklets, infographics, information webinars, and a video invitation.
- *Keep regular contact:* Some participating school systems established a school co-ordinator to liaise with the national centre and collect materials (e.g., consent forms, questionnaires). In Shanghai (China), regular communication with schools was enabled through instant online messaging applications, which allowed for feedback and support immediately.
- *Give something back to participants:* Additional to formal acknowledgement, participating teachers earned credits in Madrid (Spain), received video-based feedback in Japan, and their schools were offered financial incentives in Colombia, England and Shanghai.
- *Obtain endorsements:* The involvement of key stakeholders (e.g. teacher unions, associations of Mathematics teachers) was sought to increase the trust and relevance of the Study. For instance, England, included representatives from teacher unions in the steering committee for the project to obtain support and feedback from them.

Consent from participating teachers and students

37. The Study adheres to strict privacy and human protection regulations that require teachers and their students to provide consent before participating. The Study was carried out in a class from a sampled school only if the teacher and a sufficient number of students

consented to be video-recorded for this study. The minimum class size is 15 students. Participation by students in sampled classes required written consent from the students' parents. If more than 50% of parents of students in the class failed to provide consent to participate in any one of the Study instruments or the class size dropped below 15 students for any instrument, then the class was dropped from the sample and a replacement teacher in the school was invited to participate.

Conceptualising teaching

38. Virtually all policy-makers, researchers, and professionals agree that the quality of teaching matters. Yet, there is no common agreement on what specifically constitutes quality teaching. Conceptualising teaching was necessary for the TALIS Video Study in order to develop instruments to measure it. The Study addressed this challenge by reviewing existing conceptualisations of teaching, defining instruments that are agnostic and can capture a range of teaching approaches, and refining them on the basis of actual evidence collected in the pilot.

Review of existing conceptualisations of teaching quality

39. To develop a shared understanding of teaching quality, the Study built upon national standards and international research on teaching. The goal was to identify common aspects of teaching from across the world through a collaborative process. The conceptualisation of good teaching drew on the following sources:

- *Participating countries' own conceptualizations of teaching*: Each participating school system provided an overview of the country's conceptualization of teaching quality, its teaching standards, and research on good teaching and observation protocols. They were asked to identify ten or fewer characteristics of quality teaching.
- *Research literature on teaching quality*: Existing meta-analyses and articles on teaching quality were reviewed, and available observation and artefacts protocols were reviewed to identify aspects that have been successfully measured before.
- *Conceptual Frameworks:* Other relevant OECD conceptual frameworks such as TALIS 2018 and PISA (Table 2) to ensure alignment and coherence with other OECD work.

40. Overall, there was good alignment amongst the reviewed sources on the substantive aspects of teaching quality. Most differences related to what teaching aspects were highlighted, how these were grouped, how much abstract or detail was provided for each of them, and how subject-specific they were.

A careful approach to defining teaching

41. The conceptualisation of teaching was developed in consideration of several premises:

- *No one effective way of teaching*. The goal was not to identify *the* globally most effective way of teaching. The approach adopted was rather comprehensive, recognising that a) teaching does have multiple goals, b) different practices and features of teaching are relevant for different goals, and c) these relationships as well as pedagogical norms may vary between cultures.

- *Focus on "how to teach"*. Research on teaching can be categorized into two types, research focused on "what to teach" and research focused on "how to teach". The literature that was reviewed has been that of the "how to teach" strand, even if completely removing "what to teach" is impossible.
- *Teaching during classroom instruction*. The conceptualisation of teaching quality focused on what occurs during classroom instruction. As a result, the following aspects of the school life were not included: a) the surroundings (e.g., the school climate, cooperation with teachers or parents, country-level norms and values), b) reflection of teaching, and c) summative assessment/assessment of learning (as opposed to formative assessment which aims at adapting instruction to students and therefore was included).
- *Looking beyond what the teacher does*. While the Study is primarily interested in what the teacher does, to a large extent teaching is a combination of teacher and student actions. Some classroom practices do focus solely on the content (e.g., ensuring correct and coherent treatment of content), but practices can also related to the students (e.g., supporting social relationships among and between teacher and students), or be about the students in interaction with the content (e.g., encouraging cognitive student engagement). Therefore, the practices considered in the conceptualisation of teaching quality reflected actions taken by the teacher and by the students, avoiding a completely one-sided view of practice.

The six domains of teaching identified

42. The conceptualisation of teaching in the TALIS Video Study breaks down teaching quality into six "Domains of Teaching Practice" (Figure 3). This conceptualisation has guided all the instrument development of the Study – including questionnaires, observations, or artefacts. The six domains are:

- *Classroom Management:* The processes that ensure lessons run smoothly and efficiently so that the teacher and students spend the maximum time on academic and social emotional learning.
- *Social-Emotional Support:* There is a supportive learning environment characterized by a positive climate, with students willing to take risks in the classroom and being challenged on an intellectual and sometimes emotional level.
- *Quality of Subject Matter:* Content and tasks are clear and accurate, and students and teachers make explicit connections between the subject matter's ideas, procedures, perspectives, representations or equations that are clear and appropriate.
- *Discourse:* There are extended conversations between and among the teacher and students where students do a good deal of the talking. This includes questioning that requires students to engage in a range of levels of cognitive reasoning.
- *Student Cognitive Engagement:* The students engage in analysis, creation, or evaluation work that is cognitively rich and requires thoughtfulness.
- *Assessment of and Responses to Student Understanding:* Teachers elicit students' understanding, assess it, and respond to it by aligning their instruction to student thinking.

Figure 3. Six domains of teaching of the TALIS Video Study

43. As can be seen from Table 2, the majority of dimensions of teaching quality in the TVS can be aligned with the dimensions previously distinguished in the conceptualisation, in TALIS 2018, and in PISA.

Table 2. TVS Conceptualisation of Teaching Quality

Domain in instruments (observation protocol, artefact codes, questionnaires)	Dimension in jurisdiction's integrated conceptualization	TALIS 2018	PISA
Quality of Subject Matter	Content coverage	-	OTL
	Content-related structure	Clarity	
	Practice/Proceduralisation	-	
Discourse	Cognitive demand	Cognitive Activation	
Students Cognitive Engagement			
Responsiveness and Assessment	Adaptive teaching	Assessment and Feedback	
		Assessment Practices	Adaptivity, Perceived Feedback
Socio-emotional support	Socio-emotional support	Teacher Support	
Classroom Management	Classroom management	Disciplinary Climate	

Measuring teaching

44. The conceptualisation of teaching informed the development of the protocols to measure teaching. The teaching quality domains identified in the conceptualisation are generic dimensions of teaching quality that reach across subjects and content. The domains provide the structure and general substance for the relevant aspects of teaching to be measured. These domains were thus turned into specific classroom behaviours which can be measured.

Observation coding protocol

45. The words and phrases that comprise the observation and artefact codes take complex classroom behaviours – visible or invisible – and codify them. Most behaviours are clearly visible, such as when students work in groups, but others are harder to perceive unless attention is specifically focused on them. This could be features of practice such as a teacher aligning instruction to scaffold student learning. The observation codes help provide a tangible way of focusing attention on these important and yet often less overt aspects of classroom practice.

No judgement on quality

46. The observation codes make no judgement on quality despite being informed by a pre-existing conceptualisation of good teaching. The codes are agnostic about the relationship between a certain level of the code and outcomes of interest. Raters will code the domains as observed.

47. The posterior analysis of the data will reveal the varied relationships between those practices observed and measured student outcomes. Whether a teaching practice is "good" will only be revealed once it is related to student learning gains and other non-cognitive outcomes. Moreover, even in this case, it is unlikely that a single way of effective teaching will emerge. For example, two of the highest performing countries use different instructional approaches. In Shanghai, teachers use highly procedural instruction. In

contrast, Japanese teachers start the lesson with an open-ended question and students go through their own creative struggle to work on it.

A universal language for teaching

48. The observation protocol provides a common language to refer to teaching on the basis of specific behaviours that can be objectively observed in a classroom. It has been carefully designed to transcend national boundaries, but to still incorporate the variation that can exist between and within cultures and countries. Moreover, the observation codes are centred on evidence, and thus enable to measure the type, amount, and skill level of specific teaching practices.

49. The development of the codes was designed as a collaborative process. This was important as it would allow for the identification of pedagogical techniques that are universally applicable. To deliver on the high level of specificity on classroom behaviours required for the TVS, the development had to take into consideration how a code would apply in each specific jurisdiction, in a specific lesson or a specific set of activities within a lesson, and how specific constructs were being defined vis-a-vis the codes.

50. The observation coding protocol needed to capture the variation in teaching practice within and across the classrooms in all participating school systems. Repeatedly testing the codes on classroom videos helped focus development efforts around this goal. Nuances and discrepancies were discussed iteratively with participating school systems until there was a set of constructs that captured teaching adequately.

51. An example of the challenge of developing an international code for teaching is in capturing classroom management techniques. The practices used in British classrooms are different from those found in Japanese classrooms or in Latin American classrooms. It was important not to privilege behaviours that would only appear in some countries, such as students leading the class in a bow at the beginning and end of class, or calling out attendance at the start of class. Nevertheless, the code needed to still be able to capture such behaviours in the codes as markers of classroom management. In this specific example, this was achieved by the codes defining "routines" and asking the rater to judge the efficiency of the routine. Bowing and calling out attendance are both specific examples of routines that can be observed and judged by the rater.

Design principles

52. The codes have been designed along several key principles that aim to:

- Capture the variation within and across participating school systems.
- Maintain the same meaning and interpretation across countries, regardless of the school setting or cultural background. Therefore, a practice coded as 3 is considered a "3" in any country on any lesson by any rater at any time in the main study rating window.
- Be suitable for being captured by video.
- Allow a scalable and affordable rating process.
- Capture a broad range of different aspects of teaching.

Six domains of teaching

The final observation code divides teaching into six domains. Across these there are then eighteen components and nineteen indicators. These are shown in Table 3 below.

Table 3. Domains of teaching and respective components and indicators

Domain of Teaching	Components	Indicators
Classroom Management	Routines Monitoring Disruptions	Time on task Activity structure and frequency Time of lesson
Social-Emotional Support	Respect Encouragement and warmth Risk-taking	Persistence Requests for public sharing
Discourse	Nature of discourse Questioning Explanations	Discussion opportunities
Quality of Subject Matter	Explicit connections Explicit patterns and generalizations Clarity	Explicitness of learning goals Accuracy Real-world connections Connecting mathematical topics Mathematical summary Types of representation Organization of procedural instruction
Student Cognitive Engagement	Engagement in cognitively demanding subject matter Multiple approaches to/perspectives on reasoning Understanding of subject matter procedures and processes	Metacognition Repetitive use opportunities Technology for understanding Classroom technology Student technology Software use for learning
Assessment of and Responses to Student Understanding	Eliciting student thinking Teacher feedback Aligning instruction to present student thinking	

High and low inference codes

53. Teaching and learning behaviours are tracked and coded in different ways depending on the level of judgement required from observers. Research has shown that behaviours that require little judgement from observers (low inference) are not associated to student outcomes. In contrast, behaviours that require a higher level of judgement from

observers (high inference) can be associated with student outcomes. Accordingly, two scales were defined to capture teaching from a more descriptive and quality perspective:

- Indicators scale: Descriptive scores capture whether something happened or did not (e.g., working in pairs or in a whole class structure, the use of a calculator).
- Components scale: Quality codes capture the internationally established level of quality of a specific teaching construct (e.g., the explicitness of the lesson's learning goal, the quality of questioning in a lesson).

54. Each construct (indicator or component) has associated with it a behaviour with a descriptor for each score point. It also has a video anchor, which is a brief video clip that demonstrates what the code descriptions mean. While the practices are generalisable, the specific examples provided for each construct were based on mathematics and more precisely on the focal unit.

Figure 4. Example of an indicator and component from the 'Quality of subject matter' domain

Indicator

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).

Component

Explicit connections.	**1**	**2**	**3**	**4**
Teacher or students make explicit instructional connections between any two aspects of the subject matter. Aspects include subject matter ideas, procedures, perspectives, representations, or equations.	There are no instructional connections between ideas, procedures, perspectives, representations, or equations. OR Connection(s) that are present are implicit.	There is one instructional connection between ideas, procedures, perspectives, representations, or equations. AND Connection(s) are generally explicit, but vague.	There are at least two instructional connections between ideas, procedures, perspectives, representations, or equations. AND Connection(s) are generally explicit, clear, and brief.	There are at least two instructional connections between ideas, procedures, perspectives, representations, or equations. AND Connection(s) are explicit and clear, and at least one is elaborated.

55. In the TALIS Video Study, lessons are coded only on either high or low inference codes at a time. The lesson is also broken down into different segments, which are primarily coded independently of each other. Components are coded on longer segments (16

minutes), while indicators are coded on shorter segments (7 minutes). This approach results in multiple scores that are then aggregated.

56. The approach of segmenting the coding can help yield more reliable estimates of teaching practice at the lesson or class level. Videos are a complex stimulus where observers see many different things simultaneously and in quick succession. Raters can easily feel overwhelmed, which can in turn affect the reliability of their ratings. The segmentation reduces the cognitive demand placed on raters by shortening the amount of time and number of details a rater must consider. This should increase the likelihood of accurate and reliable coding (Bejar, 2012), which has, typically, been a common challenge in observation codes (Bell et al, 2012; Kane and Staiger, 2012).

An iterative development

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

58. International experts viewed the videos and tested the codes as they were written. They rated each segment of the video, provided evidence for those ratings, and, where applicable, wrote notes on the codes themselves about how they were functioning. Discussions ensued to reach consensus on the ratings, and to agree any edits to the coding protocols.

59. Participating school systems took an active part in the development of the codes as well. They provided input on:

- Developing the understanding on the degree to which the behaviours that were codified in the codes were prevalent in their classrooms. The contextual factors at play 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 school systems 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 behaviours being measured. This included how raters would be able to find evidence to rate particular behaviours, 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 school systems.

60. Additionally, the TALIS Video Study Technical Advisory Group also reviewed the observation codes. These reviews yielded questions about the codes and provided additional ideas for solving coding issues that had been raised.

61. The development of the observation coding protocols drew on actual footage from participating school systems. Videos collected in the pilot of the TALIS Video Study were rated with various versions of the observation codes, and specific videos were selected to

be used for training/development examples, benchmark clips, as well as for certification, calibration, and validation.

Building capacity for rating

62. Under the leadership of national master raters, over 100 raters from 8 different school systems took part in the process of rating of TALIS Video Study videos. Master raters received an extensive training into the Study aimed at preparing them for the challenging job with raters, and reducing the inherent risks of the train-the-trainer model.

63. It was important that the master raters developed a deep and shared understanding of the codes so that when they oversaw their own jurisdiction's training and operational scoring, they could answer questions that arose from country specific issues not identified in the development process.

64. The preparation that master raters received included the following elements:

- *Review and Collaboration.* Master raters reviewed codes and collaboratively coded pilot and main study videos. Master raters annotated videos where they thought the master ratings were wrong or where they had questions about how the codes should apply to a specific set of behaviours. This took place in an in-person meeting in Mexico City from 14-25 May 2018.
- *Practice rating*. Master raters were asked to independently code previously master rated videos, finding errors and discrepancies in the master ratings and collaboratively clarifying the benchmarks for various score points. This process was carried out in voluntary 2 week cycles over 12 weeks which concluded with a review of the patterns in master raters' cycle work, answering questions, and discussing issues.
- *Training and certification*. There was a review of the final version of the observation codes with all master raters. This included activities and discussions to establish a similar understanding of the codes across all countries/cultures. At the end of the training session, master raters took a certification test, for which they scored two full-length quadratic equation lessons. Master raters who did not achieve the established standard of accuracy at the first try received additional support from the observation development team and then took a second certification test. This took place in a face-to-face meeting in Pittsburgh (United States) from 20 October to 3 November 2018. Because the Study utilised a train-the-trainer model, the training process mirrored the process intended for master raters to use in their own country in training their own raters.

65. Accordingly, master raters then recruited and trained raters in their country. Raters were expected to have a mathematics background, including in quadratic equations, but they were not necessarily mathematics experts. The goal was that raters were as similar as possible in the way they understood classroom interactions and would apply the scoring codes to the classroom videos. To this end, the training was designed to discipline their thinking so that the scoring scales would have the same meaning across raters within and across participating school systems as well.

Quality assurance

66. Rating is about making judgements. Raters are trained to discipline their thinking through structured note-taking and frequent references back to the rubric language. It is this

process of watching and reading, note-taking, and referencing the rating criteria that supports the creation of valid and reliable scores.

67. Risks to validity are still inevitable though. It was clear from the outset that exact agreement on the rating is not possible, because of the fact that raters are making judgements. Accordingly, the following quality checks were introduced to reduce the risks of measurement error:

- The process of developing master ratings was collaborative and shared with all master raters.
- Certification was introduced to ensure that individual raters were obtaining a certain level of accuracy in rating.
- Weekly calibration rounds in particular to monitor the accuracy of ratings. This provided evidence about the extent of rater agreement with the master ratings. To further ensure consistent use of the codes, calibration discussions with all raters were held to review this and any particular codes.
- Weekly validation rounds were included in the assignments given to raters, for both videos and artefacts. This allowed master raters to monitor raters' work.
- Segments were rated by two different raters to ensure precision.
- Internationally agreed master ratings were used to ensure common benchmarks.

Artefacts coding protocol

68. The TALIS Video Study is unique in its efforts to collect and analyse lesson artefacts to understand their role in student learning at an international scale. Artefacts can provide information about classroom practice, and their collection is simpler and less intrusive than videotapes.

Defining the artefacts to be collected

69. The artefacts collected included lesson plans, in-class and homework assignments, visual materials, textbook pages, and formative assessments and unit tests. It was not possible to collect samples of student work nor was it possible to collect annotations or marked work from teachers. While this could have provided both contextual information and reflections about each day's lesson, issues of consent and burdening participants impeded its collection.

Piloting the collection of artefacts

70. As part of the pilot, each participating jurisdiction submitted artefacts for a subset of teachers and a subset of instructional days. Between 7 and 19 teachers participated across the 8 participating school systems. Artefacts were requested from the lesson that was video recorded, and the next lesson on quadratic equations, usually occurring the next day, even though it would not be filmed.

71. In total nearly 1 000 individual instructional artefacts were collected as a part of the data collection pilot. This consisted of the following:

- Daily lesson plans (if the teacher usually prepares a written lesson plan) (15%)
- Visual materials teacher shows to students (e.g., slides, transparencies, PowerPoint images) (27%)

- Textbook pages or other written material teacher asks students to read (14%)
- Assignments, worksheets or activities teacher asks students to complete (14%)
- Homework assignment or work students take home to complete (only if it was assigned) (9%)
- Brief "formative" assessments (e.g. exit cards) or quizzes teacher asks students to complete (6%)
- Other teaching materials used during the lesson (4%)
- Unit Test (9%)

The domains of teaching to be coded

72. The artefacts code considered the domains of teaching defined in the conceptualisation of quality teaching. However, the domains of "Classroom Management" and "Socio-Emotional Support" were omitted from the artefact code as the artefact materials provide insufficient evidence to make reasonable claims about instruction. The final domains and specific components considered within them are indicated in Table 4.

Table 4. Scorable artefact components for each practice domain

Domain Name	Component Name
Descriptive information	Scorability (Sufficient Evidence For Rating) Subtopic Coverage
Quality of subject matter	Accuracy Of Materials Explicit Learning Goals Addressing Diverse Student Needs Connecting Mathematical Representations Explicit Patterns And Generalisations Real-World Connections
Discourse	Asking For Explanations
Student cognitive engagement	Using Multiple Mathematical Methods Opportunities To Practice A Skill Or Procedure Technology For Understanding
Assessment of student understanding	Encouraging Student Self-Evaluation

Coding protocol for artefacts

73. The development of an artefact coding protocol was based on the conceptualisation of quality teaching, research and country practice. Each of the identified domains of quality teaching was operationalised through a set of specific associated components. These components were not designed to cover all aspects of the domain, but rather, they are reflective of the expectations about which aspects of each domain will be observable in artefact-based evidence.

74. The development of the artefact code followed the same cycles and collaborative iteration as the video observation codes. As in the case of videos, the artefacts collected in the pilot were also instrumental for redefining the artefact codes. Based on this feedback, several components were modified and scoring rules clarified:

- "Accuracy of Materials" was simplified to reduce the burden on raters.

- The definition of "Connecting Mathematical Representations" was clarified, and representations within this was specifically defined.
- "Explicit Patterns and Generalizations" was revised to focus specifically on inductive reasoning and prediction.
- The meaning of "Practice Opportunities" was clarified

The rating process

75. The rating of artefacts is less complex than that of videos. The artefacts cannot by their nature reveal as much information as videos. Artefacts cannot capture the "whole" lesson from start to finish. The range of teaching practices that can be captured through artefacts is also more limited. Thus, ratings artefacts entail only a moderate inference on the part of coders.

76. Each component is rated on a three-point scale (with the exception of Accuracy of Materials, which is rated on a two-point scale), including anchors for low (1), medium (2) and high (3). In general, the low level for each component provides information about the presence of a specific teaching practice or intended learning opportunity. Medium and high levels are distinguished based on differences in the nature of a specific practice or intended learning opportunity.

77. For many components, the differences in the scale reflect the extent to which students have agency or authority (i.e., the extent to which students are the source of mathematical ideas), and the extent to which there is evidence that practices or learning opportunities are intended to develop conceptual understanding in addition to procedural fluency.

78. The rating for a given lesson is actually the highest score given to any of its artefacts. As individual artefacts considered separately do not necessarily reflect the entire lesson, raters were asked to consider all artefact evidence in assessing the quality of teaching for a given lesson.

79. Each rating point of the artefact code has associated a benchmark and reasoning example to assist raters in applying the codes consistently and accurately (Box 3). Research on artefact-based measurement of instructional practices (Borko et al, 2005; Gitomer et al.) has shown that the provision of exemplars at every performance level can help raters understand the components and levels and support them in applying the codes to artefacts.

Box 4. Example of artefact rating

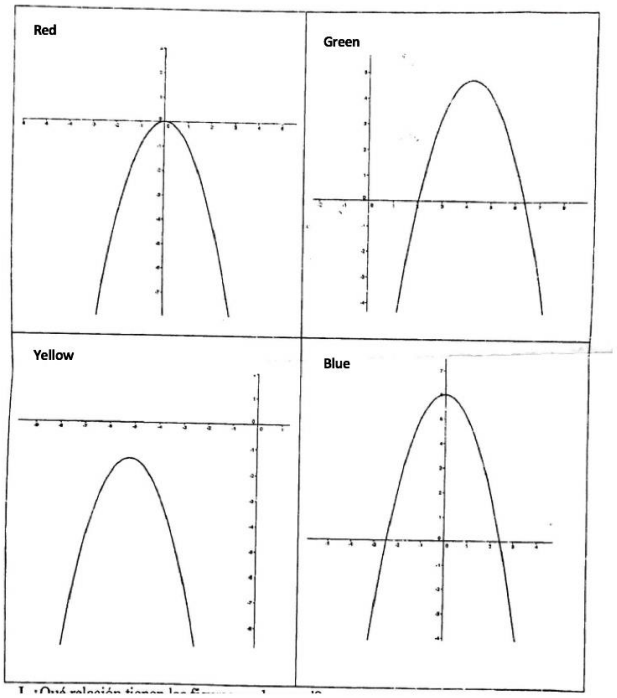
Asking for explanations measures the extent to which students are asked to explain or justify their thinking about mathematical procedures and concepts. The following examples illustrate what rating artefacts on this measure looked like:

Solve the following by factoring:

a) $3x^2 - 18x = 81$ b) $3x^2 + 6 = x(x + 13)$

The above was rated **low** because students are only asked to apply a standard routine; they are not asked how or why the method works, or why it is appropriate.

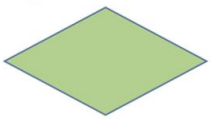
Next, 4 figures are provided. The cards have an orientation according to a side marked with a black line, which has two positions; down and up.



I. What's the relationship between the green and blue figures?
II. Is there any difference or similitude between the green and blue figures, with the red figure?
III. What's the relationship between the red and blue figures, with the green and yellow figures?
IV. What's the difference between the yellow figure and the others?
V. Explain what happens between the axes of the plot and the graphs.

This was rated **medium** because at least some questions require students to explain how to perform procedures. Students are asked here to explain a relationship, namely how the transformation shifts the graph. It is not, however, given a higher rating as it does not ask why particular procedures work or why concepts are true by having students explain how transformations are related.

4. Calculate the lengths of the diagonals of a field rhombus shaped, knowing that they differ in 6 m and the area of the field is 108 m².



a) If you assign the *x* variable to one of the diagonals, how do you represent the measurement of the other diagonal? How do you represent the area of the field?
b) Which quadratic equation solves the problem? Which method is convenient to use for solving? Why?

Finally, the above was rated **high** because students are asked to describe a procedure they used, such as how to represent the area of the field, and to provide the justification for why procedures are effective in a given solution.

Teacher and student questionnaires

80. Teachers and students were asked to complete questionnaires before and after the videotaped lessons to provide greater information about teaching and student learning (Table 5). The student questionnaires covered information on context, input, processes, outcomes of student learning, and insight on non-cognitive dispositions such as a student's motivation and interests. The teacher questionnaires, additionally, contained items that reflect the focal topic of the lessons and aspects of quality best understood through the teacher perspective.

Table 5. Content of Main Study Questionnaires

Student Questionnaires	Teacher Questionnaires
Background of Individual	
Gender, Date of birth, Migration, Parental Education, Home Possessions	Gender, Age, Formal Qualification, Teaching Qualification, Education in mathematics, Work experience
Dispositions	
Interest in mathematics, self-efficacy in mathematics, self-concept in mathematics, Instrumental motivation, Learning goal orientation, Effort and perseverance,	*Self-efficacy/Enthusiasm/Emotions teaching the target class*, Enthusiasm/Self-efficacy beliefs in general, Knowledge as it is used in Practice; Constructivist beliefs, Responsibility, Job satisfaction
Teaching	
Classroom management, Discourse, Cognitive activation, Teacher Support, Student-Teacher relations, OTL, Clarity of instruction, Focus on meaning, *Adaption of instruction*, Assessment practices	
Student-Student-Relations, Expectations, Use of learning opportunities, Homework Assignment	
School and Class Context	
Learning time	Teacher collaboration, Teacher autonomy, Factors hindering instruction, Quantity of instruction, Teaching goals
Study Context	
Reactivity videotaping, Typicality of lessons	
Perception of the Tests	Experience of being videotaped

Note: Content in italics has been included both in the respective Pre-Questionnaire (asking for general baseline information) *and* the Post-Questionnaire (referring to the focal unit). Otherwise, content is used either in the Pre- or in the Post-Questionnaire.

81. The initial selection of constructs was informed by the conceptual frameworks and questionnaires developed for TALIS (TALIS 2018 and 2013) and PISA (PISA 2015/2018 for general aspects and PISA 2012 for mathematics-related constructs) (Table 2). The rationale was to ensure alignment with other OECD surveys to support the improvement of measures of teaching in future international studies. The Video Study will be able to provide information on the validity of self-reports on classroom teaching.

82. The questionnaires were also developed in consideration of the design and focal content of the Video Study. Some constructs were adapted to capture better content-specific elements of mathematics teaching and learning in relation to quadratic equations. Also, constructs to provide additional detail on instruction were added. These included the emotions of the teacher when working with the target class, teacher knowledge as used in

practice, planning time, social desirability, and interest in mathematics. Moreover, the following constructs were largely developed from scratch: student opportunities to learn, including student self-efficacy and experience with mathematics tasks; the goals of the lesson in relation to the focal unit; the teacher's experience of being videotaped; and the teacher's history of teaching the current class.

83. Importantly, one further construct that was developed specifically for the Video Study was the "Teacher Log". This code tracked the coverage, duration and order of the subtopics that were taught within the study's focal unit of quadratic equations (Figure 5). The Teacher Log captures how the teaching of the focal unit across lessons was planned.

Figure 5. The Teacher Log captures how the focal topic was taught across lessons

Please enter date and duration for each lesson on quadratic equations. In each column corresponding to a subtopic write 0, 1 or 2.
0 = this topic was not taught today
1 = this topic was a minor focus of instruction today
2 = this topic was a major focus of instruction today.
Keep this table until it is recollected with your Post-Questionnaire.

School ID: _____ Teacher ID: _____

Date	Duration of this lesson (min)	Subtopics: 0 = not taught; 1 = minor focus; 2 = major focus									
		Handling algebraic expressions (working with brackets and terms)	Binominal formulae: a^2-b^2 or $a^2+2ab+b^2$	Introducing one form of a quadratic equation	Solving quadratic equations by ...				Discuss different cases of $ax^2+bx+c=0$ depending on values of a, b, c	Quadratic functions (definition, plotting and transforming graphs, etc.)	Real life applications
					completing the square	<factorizing>	quadratic formula X= (-b +/- SQRT(b^2-4ac)) / 2a	finding roots in a graphical representation			

84. The pilot results informed the further revision of the questionnaires to ensure that at least 75 percent of the respondents should be able to finish the questionnaires within 30 minutes (students) or 35 minutes (teachers) as well as that all questions and items must have proven measurement quality in the majority of the participating countries (e.g. reliability, no abnormality in missing responses). The revision resulted in shortening 235 items of SQA, 17 items of SQB, 225 items of TQA and 143 items of TQB. Additionally, the teacher log was also shortened and revised.

Student pre- and post- achievement test

85. Students' mathematical knowledge was tested before and after the focal topic was taught to measure student learning gains in those lessons. The pre-test was designed to measure students' general mathematics knowledge, their pre-requisite content for the focal content, and their pre-existing knowledge of this focal content. It was administered prior to the video-recording of a lesson. It essentially provided a baseline of students' mathematical knowledge of skills hypothesised to be foundational for understanding quadratic equations.

86. The post-test measured students' knowledge of the focal topic of the Video Study. It was administered within two weeks of the end of the quadratic equations unit. By narrowing the focus to quadratic equations, the post-test provided more precise estimates of knowledge and understanding of the focal unit.

87. The student test consisted of multiple-choice items. The use of constructed-response items to measure higher order thinking and creative problem solving was ruled out due to issues of reduced feasibility in the implementation and reliability of such items at an international scale.

88. The test blueprint outlined the major sub-topics of the unit on quadratic equations, and helped to determine which of these were not taught in particular countries. Each country was asked to provide approximately 10 multiple choice items that covered different aspects of the blueprint. In doing so, they were asked to consider issues of 'translatability' and to avoid items that depended on a local context that might be unfamiliar to test takers in different countries. Hence, they were asked to refrain from using, for example, the dimensions of a baseball field as a context for a question. The items were also adapted and revised to meet guidelines for style and fairness.

89. The pilot included two pre-test forms and two post-test forms of 30 items each, with the final 15 items being identical. This is consistent with current standard practice. For example, forms were assembled based on the test blueprint and items were selected so as to avoid clueing, where one item may give a hint that would help answer another. Further changes were limited to just translation issues and any necessary context adaptions (e.g., changing names or currencies).

90. The achievement test was next piloted with at least 100 students per pilot form. After data cleaning the following item statistics were calculated: classical item statistics, alpha reliabilities for each form, distractor analysis, and total score descriptive statistics. These helped consider whether items were too difficult or too easy, relatively more or less difficult in some countries than in others, or had a negative bi-serial correlation in any country. Nevertheless, due to the limited size of the item pool, a small number of items with negative bi-serial correlations in one or two countries had to be included in order to cover the content as specified by the blueprint.

91. The final form of the pre-test included 30 items while the final form of the post-test only 25. The post-test contained more items of conceptual understanding than the pre-test. As conceptual understanding items take longer time to complete than procedural knowledge ones, the overall number of items in the post-test was lower.

Analysis and reporting of findings

92. The final report will inform on the findings of the TALIS Video Study. The methodology for the analysis of the wealth of data that has been collected is explained in Box 5. The tentative outline of the report and key analytical questions to be examined in each chapter are provided below:

- **Chapter 1. What can we learn about teaching by looking directly into the classroom?** The purpose of this chapter is to describe the goals of the study, explain why understanding the relationship between teaching practice and student outcomes is important for educational improvement, and pinpoint the unique contributions of the study to the international literature on teaching practice.
- **Chapter 2. How can we understand and measure teaching practices?** The purpose of this chapter is to provide an overview of the study design, a description of how teaching was conceptualised and measured in the study, a description of the samples in each of the participating school systems as well as any issues that may impact the quality and/or comparability of the data collected.
- **Chapter 3. How do teaching practices relate to student outcomes?** The purpose of this chapter is to look at the relationship between the domains and indicators of mathematics teaching, in terms of both observations and artefacts, and the student outcome measures. The chapter will report on the student cognitive and non-

cognitive outcomes in each school system, relationships between mathematics teaching practices and student outcomes, and any existing common patterns found between school systems. It will also look on whether what is taught (opportunities to learn) has an impact on student outcomes.

- **Chapter 4. What does teaching look like?** The purpose of this chapter is to describe the teaching practices in the identified domains of mathematics and how they varied among school systems. It will explain what are common and variable patterns of mathematics teaching practice, how are teaching practices related to one another, and how are teaching practices related to characteristics of schools, classrooms, and students.
- **Chapter 5. What do opportunities to learn look like?** The purpose of this chapter is to describe the variability in the implemented curriculum across participating school systems. This chapter is important for two reasons. First, it provides background information that helps build an understanding of the differences in how variable the implemented curriculum is within and between participating school systems. Second, "opportunities to learn" has been widely discussed as key for understanding student achievement in international studies, including recent rounds of PISA.
- **Chapter 6. What teaching practices are associated with teacher characteristics and dispositions?** This chapter will explain how mathematics teaching practices vary by teacher characteristics, what common patterns are there in the relationship between teacher characteristics and mathematics teaching practice, what is the relationship between teacher disposition and teaching practice, and how are teacher and student self-reports of practice related to other measures of teaching practice.
- **Chapter 7. What are the implications on classroom practice for policy-makers and teachers?** This chapter will look into how variable is the relationship between mathematics teaching practice and student outcomes, and what does that mean for educational improvement. It will also consider what we can learn in terms of teacher allocation by experience or school contexts to make use of effective teaching practices. It will also provide thoughts for the future of large-scale observational studies of teaching, and how they can help us support teachers and improve our education systems.

93. Because any given system will be part of an international study with common instrumentation and procedures across all countries and economies, each system will be able to interpret its results in the context of the findings from the other eight participating systems. Interpreting the results in the context of multiple countries and economies may lead to fresh or deeper ideas about how to conceptualise teaching and ways to improve teaching and learning which would not be possible without the international perspective.

Box 5. Methodological approach to the analysis of TVS data

Firstly, summary statistics for the student, school and teacher background characteristics will be generated. These will then be compared to the samples of the most recent PISA and TALIS surveys from the given participating school system. Next, the relevant teaching practice scales will be produced from each data source. This involves testing for survey item functioning for each item separately by jurisdiction, and creating scores using common items. The scores for teaching practice will then be compared across different sources (i.e. observations, artefacts, etc.) using correlational analysis by school system, with and without adjustment for measurement error. The relationship between student education outcomes and teacher practices will be estimated using regression analysis, where each domain of teaching practice will be included individually or as a group. The relationship between teacher practice and teacher background/dispositions will also be conducted using regression analysis. The analysis will address a number of technical issues, such as the clustering of students within classrooms, measurement error in the pre-test score and teaching practice measures, and adjustment for multiple hypothesis tests.

References

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement, 31*(3), 2-9.
Bell, C. A., Gitomer, D. H., McCaffrey, D. M., Hamre, B., Pianta, R., & Qi, Y. (2012). An argument approach to observation protocol validity. Educational Assessment, 17(2-3), 1-26.
Borko, H., Jacobs, J., Eiteljorg, E., & Pittman, E. (2006). Video as a tool for fostering productive discussions in mathematics professional development. Teaching and Teacher Education, 24(2008), 417-436.
Borko, H., Stecher, B., & Kuffner, K. (2007). Using artifacts to characterize reform-oriented instruction: The scoop notebook and rating guide. CSE technical report 707. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The test matters: The relationship between classroom observation scores and teacher value added on multiple types of assessment. *Educational Researcher, 43*(6), 293-303.
Hill, C., Heather & Kapitula, L., & Umland, K. (2011). A Validity Argument Approach to Evaluating Teacher Value-Added Scores. American Educational Research Journal - AMER EDUC RES J. 48. 794-831.
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*.
Lipowsky, F., Rakoczy, K., Drollinger-Vetter, B., Klieme, E., Reusser, K., & Pauli, C. (2009). Quality of geometry instructions and its short-term impact on students' understanding of the Pythagorean Theorem. *Learning and Instruction, 19*(1), 527–537.
Little, O., Goe, L., & Bell, C. (2009). A practical guide to evaluating teacher effectiveness. Washington, DC: National Comprehensive Center for Teacher Quality
Martínez, J. F., Borko, H., & Stecher, B. (2012). Measuring instructional practices in middle school science using classroom artifacts. *Journal for Research in Science Teaching, 49*, 38–67.
Matsumura, L. C., Patthey-Chavez, G. G., Valdes, R., & Garnier, H. (2002). Teacher feedback, writing assignment quality, and third-grade students' revision in lower- and higher-achieving urban schools. *Elementary School Journal, 103*(1), 3-25.

Matsumura, L. C., Slater, S. C., Junker, B., Peterson, M., Boston, M., Steele, M., & Resnick, L. (2006). Measuring reading comprehension and mathematics instruction in urban middle schools: A pilot study of the Instructional Quality Assessment. (CSE Technical Report #681). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
Pianta RC, Hamre BK. Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. Educational Researcher. 2009; 38(2):109–119.
Stein, M. K., & Lane, S. (1996). Instructional tasks and the development of student capacity to think and reason: An analysis of the relationship between teaching and learning in a reform mathematics project. *Educational Research and Evaluation, 2*(1), 50–80.
Van de Vijver, Fons & he, jamis. (2014). Van de Vijver, F. J. R., & He J. (2014). Report on social desirability, midpoint and extreme responding in TALIS 2013. OECD Education Working Papers, No. 107. Paris, France: OECD Publishing.
Wayne, A. J., & Youngs, P. (2003). Teacher Characteristics and Student Achievement Gains: A review. *Review of Educational Research, 73*(1), 89-122.