



Seeing Through the Eyes of Classroom Observers: The Case of Rating Contrasted Groups of Lessons with Classroom Observation Measures

Kathleen Lynch

University of Connecticut

Abstract: Classroom observations are commonly employed to assess quality of instruction in research and practice in mathematics education. However, there is more to be learned about how sensitive classroom observation protocols are to exemplars of strong mathematics instruction, and continuous refinements to observation protocols or rating processes that may be warranted. In this study, we use the public-released mathematics videos from the Third International Mathematics and Science Study (TIMSS) to examine how classroom observers, using two contemporary classroom observation instruments, rate a set of lessons whose instructional quality is in theory expected to differ, also referred to as contrasted groups. We find that descriptively, the pattern of findings is distinct from prior studies' conclusions about the relative instructional quality reflected in the TIMSS video pool. We provide qualitative examples to illustrate the findings, and discuss implications for future research. We point to the potential value of exploring classroom observation rubrics' performance using 'contrasted groups' of lesson videos, as a tool to broaden our understanding of how observation instruments are functioning.

Keywords: Mathematics teaching; Classroom observations; TIMSS; Mathematics instruction.

DOI: <https://doi.org/10.31756/jrsmte.723>

Introduction

Classroom observations are widely used in mathematics education research to measure the impacts of research interventions on classroom instruction and student learning. In school settings, classroom observations are frequently used to allocate resources for teacher professional development and coaching. Classroom observations also play important roles in education policy: In the United States context, most U.S. states require teachers to be observed via classroom observations multiple times yearly (Walsh & Ross, 2019), and a sizeable share of large U.S. school districts connect teachers' pay to their evaluation ratings (Nittler, 2020). In the early childhood context, scores on one observation protocol, the Classroom Assessment Scoring System instrument (CLASS; La Paro et al., 2004), are employed in critical decisions about funding for federally supported early childhood centers in the United States (Office of Head Start, 2015).

Several important assumptions underpin the use of classroom observations in research and practice, including that children will learn more in classrooms where the instruction is better rated and that observation instruments can distinguish reasonably well between higher and lower quality instruction (e.g., Charalambous & Praetorius, 2018, 2022). However, while positive correlations between classroom observations and test scores have been reported in numerous studies (e.g., Kane & Staiger, 2012; Mashburn et al., 2008), several recent large-scale studies have found mostly null relationships between scores on influential classroom observation instruments such as the CLASS and students' learning outcomes (Guerrero-Rosada et al., 2021; McDoniel et al., 2022; Perlman et al., 2016). Prior research at the K-12 level has documented that the accuracy and reliability of observation scores are often too weak

to support their use in high-stakes decisions (Bell et al., 2019; see also White, 2018) and are especially low when rating instructional content, compared to lower inference items measuring classroom environment or organization (Bell et al., 2012; 2015).

Such findings raise questions about potential modifications that may benefit future iterations of classroom observation measures. They also highlight the importance of research examining validity evidence for common uses of classroom observation instruments. Validity arguments typically call for various types of validity evidence to collectively support the degree of confidence one can have in using a given assessment score interpretation for a given purpose (AERA/APA/NCME, 2014; Kane, 1992). Among these, discriminant/contrasted groups, or known groups, validity evidence reflects the premise that groups known to differ in their levels of a specific construct should be expected to have differing mean scores on an instrument that intends to measure that construct (e.g., Berk, 1976; Cronbach & Meehl, 1955). In other words, an instrument designer can show evidence of discriminant validity by documenting that the instrument produces different scores for groups that should in theory be expected to garner different scores on the measure (Centers for Medicare and Medicaid Services Measures Management System, 2023; see also Cizek & Bunch, 2007).

In this study, we provide an example of the potential benefits of examining discriminant groups as a component of validity evidence for use in the continual refinement of mathematics classroom observation protocols. Specifically, we use the public-released mathematics videos from the TIMSS Video Study to address the following question: How do classroom observers, using two contemporary classroom observation instruments, evaluate the instructional quality of lessons chosen to illustrate the key features of sample mathematics instruction in the United States and Japan, as reflected in the TIMSS video study? We use observation scores assigned to the TIMSS public-access lessons from the United States and Japan as an illustrative case of how considering examples of instruction from discriminant/contrasted groups (here, the groups of public-access lessons from the USA and Japan) may point toward future areas where observation instruments' rating processes could be usefully updated and refined. As mathematics education research and practice continually evolve, discriminant validity evidence can provide a useful tool for researchers as they iteratively assess and fine-tune classroom observation instruments and processes to align with shifting uses and contexts.

Literature Review

Classroom Observation Protocols in Mathematics Education Research

The use of classroom observation tools has long been critical to mathematics education research and practice (e.g., Bostic et al., 2021; Brophy, 1986; Charalambous & Praetorius, 2018, 2022; Kane & Staiger, 2012; Klette & Blikstad-Balas, 2018; White, 2018; White & Klette, 2024). In the 1970s, a stream of 'process-product' research aimed to catalogue observable facets of teachers' instruction, such as minutes of wait time and the frequency and difficulty of academic questioning, and correlate these features with students' test scores (e.g., Brophy, 1986). Other

researchers have used observers' overall judgments to generate holistic ratings of instructional quality in mathematics classrooms, for example by providing an overall rating from poor to excellent for the extent to which classroom activities were consistent with those recommended in the NCTM *Standards* (Schoen et al., 2003).

A more recent, and burgeoning, stream of research has developed a range of classroom observation protocols that aim to capture multiple facets of instructional quality, either for general purposes or to serve specific research project aims (see e.g., Bell et al., 2019; Charalambous & Praetorius, 2018, 2022; Klette & Blikstad-Balas, 2018). A recent review (Bostic et al., 2021) identified over one hundred peer-reviewed mathematics education journal articles that employed classroom observations published between 2000 and 2015. However, a relatively small proportion of these used formal observation protocols, and only a subset of these discussed validity evidence for the classroom observation protocols utilized. The authors identified seven protocols that were used in more than one article and for which validity evidence was mentioned, two of which were CLASS and the Mathematical Quality of Instruction instrument (MQI; Hill et al., 2008). Aside from some general links, the protocols identified were found to assess disparate constructs. Bostic et al. (2021) concluded that on the one hand, differences in the constructs examined could be expected given the different purposes for which different observation protocols were designed. On the other hand, according to Bostic et al. (2021), disparities in the constructs assessed may indicate that “as a field, mathematics education researchers do not agree on, or at least are still trying to determine, what factors support high-quality mathematics instruction” (p. 19). In response to these complexities, researchers have productively compared the application of various observation protocols to a common set of three U.S. lesson videos (Charalambous & Praetorius, 2018), and argued for the development of common classroom observation protocols that synthesize key elements from various tools (Charalambous & Praetorius, 2022). It is worth noting, however, that many researchers and policymakers have converged upon a three-dimensional conceptualization of instructional quality (i.e., classroom management, student support, and cognitive activation; Praetorius et al., 2018; 2020) as their accepted model, and in particular this vision is dominant in the European context (e.g., Hofkens et al., 2023). This conceptualization substantially overlaps with the CLASS framework domains (i.e., classroom organization, emotional support, and instructional support; Pianta & Hamre, 2009; see also Charalambous & Praetorius, 2018). Thus the influence of this framework is quite substantial.

Considering Measurement and Validation Efforts for Classroom Observation Protocols

Classroom environments and instructional interactions are highly complex; as such, evaluating classroom instructional quality presents substantial measurement challenges. Validation processes begin by enunciating the expected purpose(s) for which a classroom observation instrument will be used (AERA/APA/NCME, 2014). However, examining twelve classroom observation protocols featured in a special issue of the journal *ZDM*, Charalambous and Praetorius (2018) found that the evidence to support a set of propositions aimed at aiding validation efforts for classroom observation protocols was quite variable. For example, most protocols presented evidence relating to predictive validity, or the notion that children in classrooms that accrue higher classroom

observation protocol scores will garner stronger learning improvements. However, this evidence showed “large variability,” and “[i]n several cases, the evidence generated suggests a small or even negligible relation between instructional quality and student learning” (p. 543). Charalambous and Praetorius (2018) also found that the classroom observation protocols they considered had particularly limited validity evidence pertaining to the degree to which teaching quality characteristics could be captured with observations, and whether adequate scales and score points could be constructed to encapsulate these features. The authors noted that expert dialogue and data analysis pertaining to these propositions were needed, but “such endeavors have been largely absent from the validation processes of all frameworks” (p. 542-543).

Contributions of the TIMSS Video Study Toward Developing Common Portraits of Instructional Quality in Mathematics

The TIMSS 1999 Video Study is a seminal research study in mathematics education. In the overall study, researchers videotaped lessons from a nationally representative sample of mathematics classrooms at the grade 8 level for seven countries/regions: the United States, Hong Kong SAR,¹ Japan,² Czech Republic, Netherlands, Switzerland, and Australia. Because in the original study, researchers sought to understand the extent to which eighth grade teachers in countries with high achievement on international assessments taught mathematics in similar ways, countries were sampled based on high mean performance on the TIMSS mathematics assessment (Hiebert, Gallimore, Garnier, et al., 2003; for full details of the study procedures, see Hiebert, Gallimore, Garnier, et al., 2003). Using data from the videotaped lessons, TIMSS researchers analyzed aspects of the mathematical problems; their implementation; and the quality of the mathematics portrayed.

Among the TIMSS 1999 Video Study’s most influential conclusions was that relative to other TIMSS countries’ sample lessons, the U.S. sample lessons were uniquely focused on low-level execution of procedures and shallow treatment of mathematical problems and ideas. Meanwhile, the TIMSS video lessons from Japan have received research attention for displaying strong pedagogical elements, including more time allocated to learning and practicing new content, higher problem complexity, and greater emphasis on proof as compared to the U.S. sample lessons (Geist, 2000; Leung, 2005). According to Hiebert et al. (2005), within the TIMSS video data, “the United States displayed a unique system of teaching, not because of any particular feature but because of a constellation of features that reinforced attention to lower-level mathematics skills” (p. 111). Researchers further concluded that “teachers in higher achieving countries appeared able to implement mathematical problems in a deeper way, using problems as a means of engaging students with core mathematical concepts. Teachers in the United States, in contrast, tended to reduce all problems to sets of procedures that required students only to execute routine steps” (Santagata et al., 2010, pp. 2-3).

¹ Per the TIMSS study authors, Hong Kong SAR is described as a country in the TIMSS research, despite being a special administrative region of China, for convenience’s sake.

² Japanese video recordings were collected in 1995 and re-analyzed for the 1999 study.

The Present Study

In this article, we posit that applying classroom observation protocols to discriminant groups, or benchmark video exemplars of instruction whose qualities have been widely discussed and contrasted in the literature, can yield important insights into how classroom observation protocols are operating. Specifically, this approach can shed light on whether the protocols are picking up variations in instructional quality that they should in theory be able to distinguish. To the extent that they are not, this can illuminate areas where either classroom observation raters may benefit from further training, or protocol items may warrant expansion or adjustment. Given the critical importance of classroom observation instruments to mathematics education research, and the just-described limitations of the validity evidence base for many classroom observation protocols (Charalambous & Praetorius, 2018), building our collective toolkit of approaches to continually improve classroom observation instruments is important for the many researchers and practitioners who use classroom observations. There are few instances of video data approximating discriminant or ‘known groups’ in the domain of mathematics instructional quality; the TIMSS video data are among the few, making them among the best-suited for this purpose. We use trained, certified raters for two prominent classroom observation protocols to generate scores, and provide qualitative examples to illustrate the findings. We conclude by discussing fruitful pathways for future research.

Method

Lessons

The current study used the mathematics lessons from the United States and Japan drawn from the set of public use video lessons (hereafter, the public-access [PA] videos) captured for the TIMSS video studies. According to the TIMSS video study public documentation, “the videos provide a concrete basis for interpreting the quantitative findings of the TIMSS 1999 Video Study” (Stigler, n.d.). There are 28 PA mathematics lessons in total, four from each of the seven TIMSS video study countries. For the current study, we used the eight PA mathematics lessons from the United States and Japan (four per country) as the focal lessons for the analysis. The procedure for compiling videos for public use consideration was as follows. For the USA, added lessons beyond those initially filmed for the TIMSS video study analysis were videotaped for public release. For Japan, public-use video lessons from the TIMSS 1995 Video Study were used. TIMSS research country teams then purposely selected the PA videos (four per country) using a systematic approach.³ Importantly, the PA video lessons are not claimed to be representative of each nation’s mathematics teaching as a whole. Rather, as reported, TIMSS researchers selected

³ Specifically, the TIMSS documentation states the following: “It was considered very important that lesson videos to be publicly released reflect as much as possible the kinds [of] teaching that were seen in the study sample. Selection of the tapes was conducted using a consistent approach... They followed a standard set of guidelines to focus their reviews and identified the lessons that would be most illustrative of the key characteristics of teaching in their country as represented in the TIMSS 1999 Video Study sample” (Stigler, n.d.).

PA lesson videos for each country that were reflective of the key characteristics of teaching seen in each country's TIMSS 1999 Video Study sample lessons.

Instruments

For the current analysis, we used two widely used classroom observation instruments, CLASS and the Mathematical Quality of Instruction (MQI). One, the CLASS, is content-generic; it is designed to be agnostic to the subject matter of the academic content being taught in observed classrooms (La Paro et al., 2004). The second, MQI, is mathematics-specific; it is specifically designed to capture instructional quality in mathematics classrooms (Hill et al., 2008). As we discuss below, the two protocols measure overlapping but distinct constructs, have distinct procedures for rater training, and are scored on different scales; as such, scores on these different protocols are not directly comparable. Our aim is not to compare the operation of these two particular protocols to one another, but rather to illustrate how scoring a set of lessons drawn from discriminant groups with classroom observation protocols (here, one content-generic and one content-specific) can be a useful tool for understanding how observation protocols are working. These two instruments both have over a decade of research use, have been adapted and studied in both U.S. and non-U.S. contexts, and have documented evidence of various components of reliability and validity in numerous studies (e.g., Delaney, 2012; Kane & Staiger, 2012; Leyva et al., 2015; Pakarinen et al., 2010). It is common practice in the research literature for raters to apply each of these instruments to videotaped classroom lessons (e.g., Kane & Staiger, 2012; Leyva et al., 2015).

Content-Generic Instrument (CLASS)

The CLASS is a classroom observation instrument intended to measure instructional quality across all subject matter domains (La Paro et al., 2004).⁴ It is in widespread use in both research and classroom practice. The instrument includes codes capturing four dimensions of instructional interactions. For the current analysis, we made an a priori determination to focus on the *Instructional Support* domain of the protocol, because this domain most closely captures the academic content of a lesson, which was the primary construct of interest in the TIMSS Video Study (Hiebert et al., 2003). The TIMSS Video Study was not centrally interested in the emotional support provided to students in classrooms, nor generic aspects of student engagement and classroom organization, which comprise the other CLASS domains. The CLASS *Instructional Support* dimensions include content understanding; analysis and problem solving; quality of feedback; and instructional dialogue. This dimension is most similar in focus to the Mathematical Quality of Instruction (MQI); however, it measures instructional support irrespective of subject matter, whereas the MQI is mathematics-specific.

⁴ Videos were rated using the UE (upper elementary) version of the instrument. A secondary version of the CLASS tool has since also been disseminated; the UE version was used in the current study due to the stronger validity evidence base for it at the time of scoring. The two versions are closely aligned and include the same dimensions.

The remaining CLASS domains measure *Emotional Support*, operationalized with codes for classroom climate, the degree of teacher-demonstrated sensitivity, and the extent to which the teacher displayed regard for the perspectives of children; *Classroom Organization*, which indexes the extent to which behavior in the classroom is well-managed, classroom productivity, and instructional formats (e.g., providing organizers, using a variety of materials, showing teacher interest in student work); and *Student Engagement*, which indexes students' focus on the learning activity (for full details, see Pianta et al., 2012). We present these domain scores for completeness and to provide context for how the raters perceived the lessons in these areas. Each dimension is comprised of 1-4 items coded on a scale from 1 to 7 (low-high). Lessons are scored in 15-minute segments.

Mathematics-Specific Instrument (MQI)

The MQI instrument was designed to capture mathematics subject-specific elements of classroom instructional quality (Hill et al., 2008). It has been used in numerous research studies, including Measures of Effective Teaching (Kane & Staiger, 2012), and in coaching interventions. The MQI measures four primary dimensions: (1) *Richness of the Mathematics ('Richness')*, operationalized using codes capturing mathematical explanations, linking and connections, developing mathematical generalizations, the use of multiple solution methods in instruction, and the quality of the mathematical language employed in instruction; (2) *Working with Students and Mathematics (WWS)*, operationalized as the strength with which teachers remediate student mistakes and challenges, and respond to children's mathematically-related verbal or written productions; (3) *Errors and Imprecision ('Errors')*, which indexes major mathematical errors in instruction, imprecise mathematical notations or language, and unclear instruction; and (4) *Student Participation in Meaning-Making and Reasoning (SPMMR)*, operationalized as the degree to which children provide explanations, engage in questioning and reasoning pertaining to mathematics, and the cognitive activation of the tasks that children engage in. Each dimension is comprised of 2-5 items coded on a scale from 1 to 3 corresponding to low, mid, and high. Raters scored lessons in 7.5-minute segments. Raters also assigned a holistic lesson-level score for 'Whole-Lesson MQI' on a scale from 1 to 5 (low-high) (for details, see Hill et al., 2012).

Procedure

Instrument Scoring and Analytic Approach

A group of trained, certified raters ($N = 16$) for the CLASS and MQI instruments, respectively, rated the video lessons. Raters for the CLASS and MQI followed two separate training procedures and rating protocols, as set by the respective developers of each protocol. Our goal was not that the raters for the TIMSS and the MQI would have identical training. Rather, our goal was that the raters would have the official training recommended by each of the respective instrument developers for their own protocols. MQI raters ($N = 8$) were recruited from the MQI rater pool, which was primarily comprised of mathematics education graduate students and current and former mathematics teachers at the secondary school and college levels. Raters were initially recruited via emails to colleagues in mathematics education departments and postings to mathematics education list-serves. MQI raters

were required to complete the MQI observer training and pass the MQI rater certification exam, both officiated by the MQI instrument developers, in order to be certified and hired as a rater. The MQI training was comprised of a series of web-based modules, designed to cover the equivalent of two full days of training content, which contained both instruction about the instrument, and online practice videos in which prospective raters practiced scoring video segments and received automated feedback about their ratings (Hill et al., 2012). CLASS raters ($N = 8$) were primarily contacts provided to the project from the CLASS instrument developer. Most were graduate or undergraduate university students. All CLASS raters completed the CLASS instrument's official observer training, which comprised a two-day in-person workshop led by a trainer from the instrument developer, and passed the CLASS rater certification exam, as prerequisites before beginning rating. Raters were randomly assigned to video lessons. Rater assignments were evenly balanced such that each rater independently viewed and scored 7 videos, resulting in a total of 112 sets of full-video scores for the full set of PA videos from the seven TIMSS countries (i.e., 4 sets of scores per video). Of these, for the current study, we used the 32 sets of scores for the lessons from the USA and Japan (i.e., two sets of MQI scores and two sets of CLASS scores for each of the eight lessons from the USA and Japan). In other words, each of the eight PA lessons from the U.S. and Japan was scored a total of four times, twice by CLASS raters and twice by MQI raters, such that each lesson was double-scored on the CLASS and also double-scored on the MQI. Raters were randomly assigned within country such that they rated one video per country.

At the completion of scoring, each PA video lesson had two sets of CLASS scores and two sets of MQI scores. We first averaged scores for codes in each dimension across lesson segments and coders to generate lesson-level scores. Lesson scores were then aggregated within country to generate country mean scores. Inter-rater agreement, computed as percentage of scores in agreement, was acceptable; following the instrument developers' recommendations, CLASS ratings were considered in agreement if within one score point (agreement rate: 0.83); for MQI, agreement constituted identical scores (agreement rate: 0.77).⁵ The MQI and the CLASS are scored on different scales (i.e., CLASS codes are scored on a scale from 1-7, while MQI codes are scored on a scale from 1-3), and the score points have different interpretations in the two protocols. As such, we do not make direct comparisons of CLASS and MQI scores, but rather we interpret MQI and CLASS scores separately.

Qualitative Illustrations

To illustrate how the comparison of contrasted groups could provide useful information for mathematics education researchers building and refining classroom observation protocols, we conducted a qualitative examination of the PA lessons from the USA and Japan. Specifically, the first author re-viewed the Japanese and U.S. PA videos and ancillary TIMSS-collected lesson materials (i.e., transcripts, lesson plans, student worksheets/tasks, and researcher

⁵ Inter-rater agreement for the instrument sub-dimensions was as follows: CLASS: Emotional Support: 0.80; Classroom Organization: 0.95; Instructional Support: 0.84; Student Engagement: 0.62; MQI: Richness: 0.78; WWS: 0.62; Errors and Imprecision: 0.95; Whole-Lesson MQI (5-point scale): 0.50; SPMR: 0.70.

and/or teacher comments) holistically, specifically attending to identified lesson segments that received similar overall scores on the MQI dimensions and similar scores on the CLASS Instructional Support domain. We present illustrative examples of contrasting lesson segments that received similar classroom observation scores to illuminate potential mechanisms contributing to the observed score patterns.⁶ For MQI, we operationalized segments with similar scores as those with similar rater-assigned segment-level ‘overall’ scores for the four major MQI dimensions (i.e., *Richness*, *WWS*, *SPMMR*, and *Errors*). For the CLASS, we focused on the *Instructional Support* domain, for the reasons described above.

Results

Descriptive Statistics

The mathematics lessons included in the PA sample had an average duration of 49.25 minutes (range: 42–53 minutes). On average, 99% of lesson segments were coded as containing half or more of classroom work time connected to mathematics, indicating a low degree of off-topic/non-mathematics-related activities in the PA lessons. The following section shows results from the CLASS and MQI ratings of each country’s TIMSS PA lessons.

Table 1

Descriptive Results of PA Lessons’ Classroom Observation Scores

	Instructional Outcomes	
	Japan Mean (SD)	United States Mean (SD)
CLASS Domains		
Instructional Support	3.97 (0.65)	4.40 (0.75)
Emotional Support	4.98 (0.53)	5.07 (0.69)
Classroom Organization	5.98 (0.21)	5.84 (0.09)
Student Engagement	6.26 (0.41)	5.83 (0.84)
MQI Dimensions		
Richness of the Mathematics	1.24 (0.12)	1.42 (0.23)
Working with Students and Mathematics	1.44 (0.25)	1.61 (0.38)
Errors and Imprecision	1.02 (0.02)	1.04 (0.04)
Student Participation in Meaning- Making and Reasoning	1.48 (0.17)	1.52 (0.36)
Whole-Lesson MQI	3.75 (0.65)	4.00 (0.71)

Note: Means reflect averages of country mean scores on each of the CLASS and MQI dimensions.

⁶ Note that we made comparisons within each instrument. We would not expect lesson segments to have similar MQI scores as CLASS scores, given that, as noted above, these instruments are scored on different scales and measure overlapping yet distinct constructs.

CLASS codes are on a scale of 1-7 (low-high). MQI dimension codes are on a scale of 1-3 (low-high). Whole-Lesson MQI is on a scale from 1-5 (low-high).

Content-Generic Instrument (CLASS)

Table 1, top panel, shows descriptive mean scores for the eighth grade public-use mathematics lessons from the United States and Japan, on the four CLASS dimensions of instructional support, emotional support, classroom organization, and student engagement, each on a scale from 1 (low) to 7 (high). For each CLASS dimension, scores of 1-2 correspond to ‘low’ performance in the domain evident in the lesson; scores of 3-5 correspond to ‘mid’-level performance, and scores of 6-7 indicate ‘high’ performance. The sample lessons generally were scored in the ‘mid’ level range for instructional support and emotional support, and in the mid-high range for classroom organization and student engagement. For *Instructional Support*, descriptively, the Japanese lessons’ mean score ($M = 3.97$) was lower than the U.S. mean ($M = 4.40$). Appendix Table S1 shows that the descriptive pattern of the U.S. lessons receiving higher average scores was observed in three of the four *Instructional Support* subdomains (*Quality of Feedback*, *Instructional Dialogue*, and *Content Understanding*), while descriptively, the Japanese lessons received higher average scores in the fourth subdomain (*Analysis and Problem Solving*). It is important to note that the TIMSS PA lesson sample was not powered for inferential testing of lesson differences by country. The differences observed are not statistically significant (Mann-Whitney U tests: *Instructional Support*: $U=5$, $p=.486$; dimension-level: *Quality of Feedback*: $U=4.5$, $p=.384$; *Instructional Dialogue*: $U=2.5$, $p=.147$; *Content Understanding*: $U=5$, $p=.457$). As such, the comparisons are descriptive and suggestive in nature only; we cannot conclude that a pattern of stronger ratings for U.S. lessons would definitively be found in a broader population of raters and analogously-chosen videos. This notwithstanding, descriptively, we also did not observe the clear pattern of stronger performance for the lessons from Japan that may have been expected based on the literature.

As seen in Table 1, for the remaining dimensions, descriptively, the Japanese lessons received higher mean scores on *Student Engagement* ($M = 6.26$) and *Classroom Organization* ($M = 5.98$). For *Emotional Support*, the United States had the higher mean score ($M = 5.07$). The observed differences are not statistically significant (Mann-Whitney U tests: *Student Engagement*: $U=11$, $p=.465$; *Classroom Organization*: $U=11$, $p=.468$; *Emotional Support*: $U=7$, $p=0.886$), again not unexpectedly given the PA sample was not powered for inferential statistical testing of cross-country differences.

Mathematics-Specific Instrument (MQI)

Table 1, bottom panel, shows the results of scoring the PA lessons using the MQI. For *Whole-Lesson MQI* (scale: 1 [low] to 5 [high]), the lessons from both the United States and Japan had mean scores between 3 and 4, corresponding to mid- to mid-to-strong whole-lesson MQI; descriptively, the U.S. lessons had a mean rating of $M = 4.0$, while the mean rating for the PA lessons from Japan was $M = 3.75$. As before, the difference favoring the US lessons is not statistically significant (Mann-Whitney U test: $U = 6.5$; $p = 0.766$); however, descriptively, the scores

are consistent with a scenario in which we did not observe the clear pattern of stronger ratings for the lessons from Japan that may have been expected.

At the MQI dimension level, with each code on a scale from 1 to 3 (low-high), both countries' lessons had mean scores between 1 and 2 on the MQI dimensions *Richness*, *WWS*, and *SPMMR*, corresponding to low-to-mid instructional quality on these dimensions. Descriptively, on these three dimensions, the U.S. lessons received slightly higher mean ratings than the lessons from Japan, although mean differences were not statistically significant (Mann-Whitney *U* tests: *Richness*: $U=4$, $p=.343$; *SPMMR*: $U=7.5$, $p=.99$; *WWS*: $U=7$, $p=.886$). Scores were generally low on *Errors*, indicating few mathematical errors in the lessons.

Qualitative Illustrations

Recall that one widely-reached conclusion from the TIMSS Video Study was that the U.S. lessons in the sample demonstrated relatively weak instructional affordances, while the lessons from Japan have often been cited for containing useful pedagogical elements (e.g., Geist, 2000; Leung, 2005). Given this, the current observed pattern of similar instructional quality ratings based on CLASS and MQI for the Japanese and U.S. PA lessons leads to questions about why this pattern may have occurred. As described above, to examine potential mechanisms, the first author re-viewed the Japanese and U.S. PA videos and supplementary lesson materials, attending to identified lesson segments for which the Japanese and U.S. PA videos received similar overall scores on the MQI dimensions and similar overall scores on the CLASS Instructional Support domain.

We echo others (e.g., Hiebert, Gallimore, & Stigler, 2003) in emphasizing that teaching is a complex cultural and professional practice; teachers are highly skilled professionals who make complex instructional decisions within the parameters of different culturally available scripts about teaching and education, professional training, student preparation, curricula, and school policies. The illustrations below are small excerpts of instruction, not sampled to be representative of specific teachers' practice; discussion of these excerpts is presented to illustrate the operation of the observation protocols, and does not characterize the instruction of specific teachers.

To preview the findings, we found that despite receiving overall similar scores on the classroom observation instruments, some PA lesson segments from Japan and the United States appeared to differ qualitatively on elements likely related to students' opportunities to learn mathematics. Specifically, in some cases, raters appeared to under-credit Japanese PA lesson segments that contained useful pedagogical elements, potentially because these elements must be inferred rather than directly observed or because raters were less accustomed to observing higher level exemplars of these practices in their training and rating practice.

We provide three examples illustrating this phenomenon below. Here, we focus on examples at the MQI (7.5-minute segment) level. With the CLASS scored in 15-minute segments, the results are not directly comparable; however,

for the 15-minute segment that contained each example, the segments received similar total points to one another on the CLASS Instructional Support dimension (see Table 2).

The first and second examples are two ‘lesson launch’ segments—one from the United States and one from Japan—that received similar scores on the MQI. The third example is the concluding segment of a U.S. lesson. This segment also received similar mean MQI scores as the lesson launch examples. We elaborate on these illustrative examples in the following section.

Example Segment #1: U.S. Sample Lesson Launch: Homework Checking and Warm-up Problems

The first example comprises the lesson launch, or the first 7.5-minute segment, of a 53-minute U.S. eighth grade lesson which was an introduction to writing variable expressions. To place this segment in context, the overall lesson proceeded as follows (LessonLab, n.d.-e). Students first spent 10 minutes working independently at their seats on warm-up problems written on the whiteboard, while the teacher circulated around the classroom checking homework. The class then spent 10 minutes reviewing the answers to the warm-up problems and homework, in most cases providing the answers without discussion of how the problems were solved. The teacher then spent approximately 24 minutes introducing new material to the class and leading the class in practicing with the new material. Students spent 5 minutes working individually on their homework, followed by 4 minutes playing a whole-class game.

Table 2 (Example Segment #1) shows an overview of this lesson launch segment (top panel) and the observation scores it received (bottom panel). Both raters assigned this clip ‘1’ scores on the MQI instructional quality dimensions (i.e., the overall codes and subcodes for the dimensions of *Richness*, *WWS*, and *SPMMR*), indicating low levels of mathematical quality of instruction evident in this clip. For the 15-minute CLASS segment that contained this clip, the segment received mean scores in the low and mid ranges on the CLASS Instructional Support dimension codes.

Regarding the warm-up activity in this segment (see Table 2, Example Segment #1, Board Illustration), the teacher described it as follows in her Teacher Comments on the lesson:

“The students do a daily “warm-up” activity... They are usually basic problems that are not related to each other in operation or subject. I find that my students need constant practice with their basic skills... While the students are working on the warm-up, I go around to each one and check their homework from the previous night. I check that it’s the correct assignment, that it’s complete, and that they showed their work, not just answers” (LessonLab, n.d.-f).

Table 2

Illustrative Lesson Segments, and Associated Classroom Observation Protocol Scores


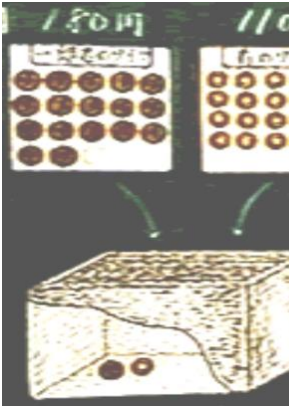
	Example Segment #1: US Sample Lesson, Launch Segment		Example Segment #2: Japan Sample Lesson, Launch Segment		Example Segment #3: US Sample Lesson, Wrapup Segment	
Primary Activity	Homework Check and Warm-up Problems		Temple Donations Problem		Homework Questions and '24' Game	
Tasks	Board Illustration ¹	Lesson Launch Task	Board Illustration ²	Lesson Launch Task	Example Homework Tasks	'24' Game
	<p>1. Is this a right triangle? </p> <p>2. $(13.0013 - 3.313) + (3.01 \times .3)$</p> <p>3. $5\frac{1}{5} - 3\frac{2}{3}$</p> <hr/> <p>4. $1.2\sqrt{4.806}$</p> <p>5. $2\frac{4}{9} + 2\frac{1}{10}$</p> <p>6. 4, 8, 2, 6 (i.e., use the numbers in 3 equations, with the last</p> <p>7. 3, 7, 9, 4 (i.e., use the numbers in 3 equations, with the last</p>	<p>Students are asked to complete the warm-up problems written on the board, while the teacher circulates the room checking for homework completion.</p>		<p>“It has been one month since Ichiro's mother has entered the hospital. He has decided to say a prayer with his smaller brother at a local temple every morning so that she will be well soon. There are 18 10-yen coins in Ichiro's wallet and just 22 five-yen coins in his smaller brother's wallet. They have decided every time to take one coin from each of them, and put them in the offertory box, and continue their prayers until either wallet becomes empty. One day after they were done with their prayers, when they looked into each other's wallets, the smaller brother's amount of money was greater than Ichiro's. How many days has it been since they started praying?”</p>	<p>Write a variable expression for each phrase.</p> <p>1. 5 less than a number.</p> <p>2. 15 more than the absolute value of a number.</p>	<p>According to the lesson graph, in this game, “students choose four numbers, and their classmates have to use them in three equations with the last equation equal to 24.”</p>

Table 2 continued

Observation Scores			
	Example Segment #1: US Sample Lesson, Launch Segment	Example Segment #2: Japan Sample Lesson, Launch Segment	Example Segment #3: US Sample Lesson, Wrapup Segment
MQI overall scores (scale: 1-3) ³			
	Mean	Mean	Mean
Richness of the Mathematics	1	1	1
Working with Students and Mathematics	1	1	1.5
Errors and Imprecision	1	1	1
SPMMR	1	1.5	1.5
CLASS Instructional Support scores (scale: 1-7) ⁴			
	Mean	Mean	Mean
Content Understanding	3	4	3.5
Analysis and Problem Solving	2	3	2.5
Quality of Feedback	3.5	3	3.5
Instructional Dialogue	4	3	4
Total points	12.5	13	13.5

Notes: ¹Board illustration from LessonLab. (n.d.-e). ²Board illustration from LessonLab. (n.d.-a). ³MQI scores are assigned to 7.5-minute segments, on a scale from 1 to 3 corresponding to low to high. ⁴CLASS scores are assigned to 15-minute segments, on a scale from 1 to 7 corresponding to low to high. Scores shown are scores for the 15-minute segment that contains the clip used in MQI scoring above.

The following transcript excerpt (LessonLab, n.d.-g) illustrates the classroom interactions during this segment:

- Teacher:* Quietly take out paper to start the warm-up. Quietly have your homework out on your desk. I'm going to come around and check it please...
- Student:* Can I borrow a pencil?
- Teacher:* Let's go. Come on. Quickly. (*teacher walking around the room checking homework*) Student A, please do the warm-up. Student B? (*inaudible*) Okay. Next time do it in pencil. Student C? What about the answer here? Okay, you have to write it out, okay? Okay.

During this segment, the teacher circulated relatively quickly between students and appeared to be checking the homework primarily for completion, rather than correctness or solution strategies used. In her Teacher Comments on the lesson, the teacher did indicate that this homework checking time helped her to gauge students' needs for review or greater repetition, which could yield important information (LessonLab, n.d.-f). However, this homework-checking method of gauging students' progress was time-consuming; coupled with the student independent work on the warm-up problems, it occupied a full 10 minutes of the 53-minute lesson. The lengthy focus on, per the teacher's characterization, unconnected tasks of a basic nature contributed to an overall impression that this particular segment contained a strong review focus, followed a relatively slow pace, and offered relatively limited affordances for engagement with higher-order mathematical problem-solving. The raters' judgments that the segment rated 'Low' on all overall dimensions of the MQI seems consistent with this conclusion.

Example Segment #2: Japan Sample Lesson Launch: Temple Donations Problem

For the purposes of comparison, we next describe an example from one of the TIMSS PA lessons collected from Japan. Like the prior U.S. example, this example comprised the lesson launch, or the first 7.5-minute segment, of an eighth-grade mathematics lesson.

This segment was drawn from a 54-minute lesson which was an introduction to inequalities and their characteristics. The full lesson proceeded as follows (LessonLab, n.d.-a). The teacher first spent 4 minutes reading a problem out loud to the class and illustrating the problem using physical manipulatives and the blackboard. Students then spent 13 minutes working on the problem on their own, while the teacher circulated through the classroom, talking with various students about their work. The class then spent 24 minutes in whole-class work, during which the teacher had five students identified during the independent work period present their solution strategies for the class on the blackboard. In the final 13 minutes of the lesson, the teacher presented a second problem to the class, then walked through the solution and its meaning with the students in a whole-class work segment.

Like the previously discussed segment, both raters assigned this segment scores of '1' (Low) for the MQI

dimensions of *Richness* and *WWS*, demarcating low levels of mathematical quality evident in the segment on these dimensions (see Table 2, Example Segment #2, bottom panel). For *SPMMR*, one rater assigned the segment a ‘2’ (Mid), while the other assigned it a ‘1’ (Low). Both raters scored all subcodes as ‘1’s (Low), with the exception that one rater gave scores of ‘2’ (Mid) for the subcodes ‘Enacted Task Cognitive Activation’ and ‘Linking and Connections.’

The teacher began the 7.5-minute segment as follows (LessonLab, n.d.-d):

Teacher: [Teacher reading the problem aloud from a paper: “It has been one month since Ichiro’s mother has entered the hospital... How many days has it been since they started praying?” (see full problem statement in Table 2, Example Segment #2, Lesson Launch Task)]
[stops reading; now talking to the class]
That’s the problem. Now, I think there are points that are a little too hard to understand with just these sentences, so- Well, I think I would like to look at a figure and check it.
Now, pay attention to the blackboard.

After reading the problem, the teacher affixed a pre-made illustration of a wallet containing rows of detachable 10-yen coin manipulatives to the blackboard. The teacher then proceeded to display the scenario posed in the task using manipulative illustrations of the wallet, coins, and offertory box (see Table 2, Example Segment #2, Board Illustration), as seen in the following transcript segment (LessonLab, n.d.-d):

Teacher: This is Ichiro’s wallet, okay?
[teacher affixes another, similar wallet illustration to the blackboard]
Okay, over here is the smaller brother’s wallet. And, also there is an offertory box there, right? Okay?
Now, first we’ll start looking at Ichiro’s, okay? If we look at the wallet there are only 10-yen coins in there, right? How much is in this in the beginning? All together. How much is in it?
Students: One hundred eighty yen.
Teacher: Yes. [teacher writes the total amount on the board in the space above the wallet]
The amount of money, well, in the beginning. Compared to the smaller brother’s. There are only five-yen coins in it, but how much is in it?
Students: One hundred and ten yen.
Teacher: Yes. One hundred and ten yen.
[teacher writes ‘110 yen’ on the board in the space above the wallet]

This means in the beginning stages, um, Ichiro has more, okay?

[teacher indicates Ichiro's wallet with his hand]

In terms of the amount of money, he has more, right?

Then, after this. Every morning from each wallet they take out one coin from each, okay?

[teacher points to Ichiro's wallet, then to the smaller brother's wallet]

They put them in this offertory box, right?

[teacher draws an arrow from Ichiro's wallet to the offertory box, then one from the smaller brother's wallet to the offertory box]

Now, first day one. Ten yen from Ichiro's wallet and five yen from over here *[while talking, teacher removes one 10-yen coin piece from Ichiro's wallet, and one 5-yen coin piece from the smaller brother's wallet, and holds them up next to each other in one hand, then slowly lowers them into the offertory box, and affixes them to the offertory box illustration]*.

Okay? Now, it goes in from here and in this offertory box.

Students: Wow. That's great. That's amazing.

Teacher: Okay. It's gets stowed away in here, okay? Is it okay?

Student: They're in there. Yes.

Teacher: Okay, then how about it at this stage? Comparing Ichiro's and his smaller brother's wallet. *[teacher gestures back and forth from the first to the second wallet]*

The contents of the wallet. Which one has more in terms of the amount of money?

Students: The older brother. Ichiro. Ichiro.

Teacher: Yes. Ichiro has more, right? Okay, then second day.

[teacher again takes one coin from each wallet, mimes lowering them into the offertory box, and affixes them to the box]

Okay. Of course right? They put in the coins in here.

How about it in this stage?

[teacher gestures back and forth from the first to the second wallet]

Students: Ichiro. Ichiro.

After this interaction, the teacher then directed the students to try solving the problem (LessonLab, n.d.-d):

Teacher: Hmm. Ichiro still has more, huh? Now, the fact is they continued things like this every day, okay? But, in the morning of some day, okay? They looked inside of each other's wallets, right? Then the smaller brother had more in the amount of money. Now, that one day was how many days since they started their praying? Now I would like for everybody

to think about that... You may want to take notes or anything else you want to do is fine, so please try doing it.

Following this public class work interaction, students began private work on the task at their desks. Meanwhile, the teacher began circulating around the room, looking at student papers and intermittently writing notes on a clipboard or talking quietly to individual students, as exemplified in the following transcript section (LessonLab, n.d.-d):

Teacher: Now, just now see? You did it up to the second day, right?
Teacher: Then, how about the third day?
Teacher: How about it?
Teacher: In the same manner try thinking, like how about the fourth day?

Later in the lesson, it became apparent that while circulating the room, the teacher had been observing which solution methods students were using. In turn, he selected students to go to the board and display their solution methods for the class, such that their approaches could be compared visually. The teacher provided the following explanation in the Teacher Comments: “As I make my instructional classroom rounds, I notate who is using what kind of solving method. When I call on the students to make presentations, I use this information” (LessonLab, n.d.-c).

The teacher further explained that he had planned to display the solutions in order of complexity and had further pre-planned the layout of the board space to facilitate the comparison of students’ multiple solution strategies.

“I had the order of the presentations start from operational methods and progress bit by bit to equation methods. I also considered the positioning of the writing on the board ahead of time so that the different methods could be easily compared afterwards. Furthermore, I had each presenting student talk about his thinking process... I would have liked to call upon a student who solved the problem with an equation using x only, but I did not come across anyone as I made my instructional classroom rounds. Therefore, I called upon a student who set up an equation using both x and y . However, because they have not learned this method yet, I intentionally omitted part of the calculation on the board” (LessonLab, n.d.-c).

There is an argument to be made that despite being rated low on the MQI dimensions of *Richness* and *WWS*, the lesson segment contained at least two important pedagogical practices endorsed as beneficial in prominent research and policy documents from U.S. and other international settings. First, the teacher posed a relatively complex problem that admitted the use of multiple solution strategies, and the students evidently solved the problem in a variety of ways (e.g., Stein & Lane, 1996; Tekkumru-Kisa et al., 2020). The teacher appeared able to understand students’ mathematical work, as evidenced by the fact that he used selected examples of this work later in the lesson

for a public display of multiple solution strategies. Specifically, in their influential framework, Smith and Stein (2018) (see also e.g., Stein et al., 2008) endorsed five pedagogical practices as generative for orchestrating productive mathematical discussions, including *anticipating*, or thinking in advance about which solution strategies students are likely to use for a problem (e.g., Lampert, 2001; Schoenfeld, 1998); *monitoring*, or circulating around the classroom during student work time, observing students' solution approaches (e.g., Fernandez & Yoshida, 2004; Lampert, 2001); *selecting*, or strategically choosing examples of students' solutions to share publicly with the class (e.g., Lampert, 2001; NCTM, 2014); *sequencing*, or positioning student's shared work in a deliberate order so as to build toward a mathematical point (e.g., Ayalon & Rubel, 2022); and *making connections between student responses*, or drawing students' attention to specific links among their solution strategies, or between their solution approaches and mathematical concepts (NCTM, 2014; Star et al., 2015).

It is apparent that prior to the lesson, the teacher engaged in the first practice, 'anticipating.' The teacher's forethought about the solution strategies students were likely to use was evident in the fact that he had pre-made labels ready for the students' methods, which he attached to the board after each student demonstration of the pertinent solution approach. Additionally, the teacher's lesson plan contained an organized table outlining the period's structure, with a column for 'Expected Student Reactions' to each problem, including five anticipated solutions to the Ichiro problem, along with corresponding teacher actions and responses to each (LessonLab, n.d.-b). For example, the teacher asked some students if they could find additional solution strategies that were faster or simpler, or by formulating a mathematical expression.

During the lesson, there is evidence that the teacher engaged in at least one of Smith and Stein's (2018) endorsed practices: 'monitoring.' According to his lesson notes, while the teacher was circulating the room, he was observing students' solution strategies. The teacher was also evidently engaged in 'selecting' as he circulated the room, as he asked five students to present their methods matching the pre-made solution labels on the board. The teacher's comments also make clear that he subsequently engaged in 'sequencing.' Per his lesson notes, he asked students to display their solutions in order from operational approaches advancing to equation-based approaches. Such a pedagogical move -- observing methods and purposefully selecting them for future public display and discussion -- is a hallmark of orchestrating strong discussions of students' mathematical solutions (Smith & Stein, 2018). It would likely not be immediately visible to a classroom observer, however, and hence unlikely to accrue points on an observation protocol.

Example Segment #3: U.S. Sample Lesson Wrap-Up: Homework Queries and a Game

The final segment of the US lesson introduced above provides a third illustration. This segment received all 'low' scores on the MQI from one rater, but was scored as 'mid' by the second rater for *Overall WWS* and *Overall SPMMR*. As a result, this segment received a higher mean score on *Overall WWS* and the same mean score on

Overall *SPMMR* compared to the Japanese launch segment described above.⁷

In the first 3.5 minutes of this segment, students worked individually at their desks on a homework worksheet, which included questions relating to the day's class material on writing variable expressions. Students raised their hands to ask questions, and the teacher circulated to their desks to respond individually. These interactions were visible to the classroom observer, earning the segment points on the rubric with one rater (albeit not with the other), despite some of the provided remediation appearing somewhat unclear. The first student who raised his hand asked about homework question #1: "Write a variable expression for the phrase '5 less than a number'" (LessonLab, n.d.-g):

- Teacher:* Student A?
- Student A:* On, uh, on number one, do you have to have- do you have to put the number before you, uh, subtraction?
- Teacher:* No. That one has to be [sounds like 'after']. So like-
- Student A:* So it would be like five subtract, uh, any number you- any letter you want to?
- Teacher:* (*addressing another student*) Thank you. (*addressing Student A*) Yes. Any letter. Yes.

Following this exchange, Student A wrote '5 - x' as the answer. Three minutes later, Student A raised his hand again. The subsequent problems on his worksheet were still blank. He again asked about question #1, as seen in the following transcript segment (LessonLab, n.d.-g):

- Teacher:* Yeah. Okay. Student A? (*walks over to Student A's desk*)
- Student A:* Yes. Right here. So, if it's five less than a number would it be five subtract X? (*shows teacher where he has written '5 - x'*)
- Teacher:* No. Because this is the number. X is the number so it has to be the other way. (*points to the 'x' Student A has written*)
- Student A:* Oh.

This time, the teacher corrected Student A's original response. However, at least some of the remediation provided to Student A was apparently unclear. Student A had to ask his question twice, after the teacher's first response appeared to lead him to an incorrect answer.

⁷ The 15-minute segment containing this clip also received more total points on CLASS 'Instructional Support' than did the Japanese launch clip, although as noted previously, the lesson segments considered in CLASS and MQI scoring differ and thus the segments are not directly comparable.

Immediately thereafter, Student A attempted to ask the teacher two more questions. The teacher responded to the first query, about how to write a variable expression for ‘15 more than the absolute value of a number.’ However, when Student A attempted to ask a third question, pertaining to the bottom part of the worksheet (“And then down –”), the student did not receive a response (LessonLab, n.d.-g):

Student A: And then down- (*indicates with his pencil a problem near the bottom of the page*)
Teacher: Wait a minute. Hang on. (*walks to the front of the room*) (*addressing the whole class*) Okay, people you may pack up. We will go over this tomorrow... We have about three minutes... Class we're going to finish by playing 24.

The teacher did not return to Student A’s question within the class period. Although the teacher may have had a variety of reasons for choosing to play a class game instead, it does seem apparent that Student A left the class with an unanswered question about his homework that he attempted to ask.

The remaining 4 minutes of the segment were dedicated to playing the game ‘24.’ According to the lesson graph, in this game, “students choose four numbers, and their classmates have to use them in three equations with the last equation equal to 24” (LessonLab, n.d.-e). The following transcript segment illustrates the game play dialogue (LessonLab, n.d.-g):

Student B: (*writing the numbers on board*) Seven, two, one, five.
Teacher: Shh. Student C. Student D.
Student D: Oh, seven minus two is five. Five times five is 25. Twenty-five minus one is 24.
Student E: Yeah.

A student kept score on the board for the teams. At least some students visibly enjoyed the game, while others made comments such as ‘that’s not fair’ and ‘cheater’ during the activity.

Holistically, there appears to be at least some case to be made that this segment did not exemplify similar or stronger instructional quality than did the example launch segment from Japan, as was suggested by the pattern of classroom observation scores. Students in the current U.S. example segment were completing a homework worksheet that was a review of previously learned skills, and a majority of the segment was dedicated to playing a game that was not clearly coherent with the overall lesson. Some remediation of student difficulties was visible, but it was arguably at least partially unclear to the student. By contrast, in the Japanese example launch segment, the teacher was not providing direct remediation during private work time, but was interpreting students’ work for a planned future display of their multiple solution strategies. The latter was less visible to an observer, but still likely valuable to the students.

Discussion

Overall, on both the dimension of the CLASS measure most proximal to the content of instruction (*Instructional Support*) and most dimensions of the mathematics-specific MQI instrument, raters assigned TIMSS PA video lessons from the United States descriptively similar mean scores to PA lessons from Japan. The disconnect between the relatively strong scores that the U.S. PA lessons received on the examined classroom observation instruments on the one hand, and the often-documented conclusion in the field that the U.S. lessons included in the overall TIMSS 1999 Video Study displayed relatively weak instructional quality on the other hand, raises questions about why this pattern of findings may have emerged. Although small score differences could be stochastic given the dataset size, we also do not see a clear descriptive pattern of stronger performance in key dimensions of instructional quality among the PA lessons from Japan that we might have expected based on prior researchers' conclusions using the original TIMSS 1999 Video Study dataset.

Based on the observed findings from the qualitative analysis, we argue that two explanations are likely. First, some facets of instructional quality that were important to the TIMSS researchers' assessments of the relative instructional affordances of the TIMSS lessons, such as pacing and coherence, are less centrally emphasized on these observation protocols, hence raters did not factor these heavily into their evaluations of the lessons. Second, there were some cases in which raters appeared to have challenges noticing especially high-quality enactment of facets of instructional quality that the classroom observation protocols were designed to measure, and distinguishing it from more mid-level enactment. We expand on each of these below. We then discuss how the current study provides an existence proof for the notion that using contrasted groups evidence, as we have done here, can be a useful tool for highlighting areas where classroom observation processes may warrant future attention and refinement.

First, one plausible explanation for the observed pattern of findings is that some features of high-quality mathematics instruction that were included in the original TIMSS rating instruments and highlighted in related TIMSS study reports (e.g., Hiebert et al., 2005; Santagata et al., 2010) may not be well-captured by the CLASS and MQI instruments. There is partial evidence to support this hypothesis. The TIMSS researchers coded lessons for several constructs not directly measured on the CLASS nor MQI, including elementary versus more advanced level of the mathematical content; time spent reviewing previously-covered material; and a lesson's overall mathematical coherence. In Hiebert et al.'s (2005) analysis of the full TIMSS 1999 Video Study dataset, U.S. lessons performed poorly on all of these indicators relative to the rest of the TIMSS country group, contributing to the researchers' overall conclusions about the U.S. lessons. Still, other constructs examined in the TIMSS study team's research do appear to overlap with CLASS and MQI constructs, including use of higher-order kinds of mathematical reasoning, such as developing mathematical justifications or generalizations; providing mathematical explanations for procedures and concepts; and working on problems in ways that maintained a focus on conceptual connections. On the MQI, codes capturing similar constructs are included in the *Richness* and *SPMMR* dimensions. The CLASS *Instructional Support* dimension includes indicators focused on the extent to which lessons stimulated in-depth

content knowledge via treatment of major organizing concepts and the development of connections among these concepts; prompted students to solve problems and explain their ideas; and included practical applications (Pianta et al., 2012); while not identical, these CLASS constructs do overlap conceptually with constructs considered in the TIMSS research.

A second likely contributing explanation is that the classroom observation raters ‘missed’ certain of the relative affordances of the TIMSS PA lessons from Japan, which have been highlighted in the literature. Specifically, raters may not be consistently distinguishing well between middling and exemplary instances of some constructs the instruments were designed to measure. This is suggested by the relatively low number of high scores that raters assigned using the MQI, and by the qualitative review of the lessons. One potential reason for this may be that raters trained and calibrated on the instruments using samples of U.S. lesson videos. Research indicates that many of the higher-level pedagogical practices which were present in the original TIMSS video study lessons are historically uncommon in U.S. classrooms (Hiebert, Gallimore, Garnier, et al., 2003; Richland, 2015), suggesting that these practices may also have been uncommon in the instrument calibration samples or enacted at lower levels of cognitive demand than in the TIMSS PA videos from Japan. Raters, less accustomed to seeing high-level examples of the behaviors the instruments aim to capture in their rating and calibration practice, may have failed to notice them when they encountered them in the PA lessons.

A third possibility we considered is that the raters simply expressed a ‘home country bias’ (e.g., Verlegh, 2007), consistent with the notion from social psychology that individuals tend to evaluate the performance of those they perceive as similar to themselves (in-groups) more favorably than that of those they perceive as dissimilar to themselves (out-groups) (Mackie & Smith, 1998). While this is possible, if this had occurred, we would have expected raters to assign U.S. PA lessons the highest scores regardless of dimension. This was not the case, as raters gave non-U.S. PA lessons the highest average scores on two of four CLASS dimensions (*Student Engagement* and *Classroom Organization*).

Taken together, the observed findings imply two conclusions. First, this study contributes to the research evidence suggesting that adjustments to rater training may be valuable to help classroom observers more effectively distinguish between high-quality versus middling facets of instruction when they see them. Since classroom observers base their scores in part on memorable training videos (Bell et al., 2015), adding more memorable training videos of mathematics segments exemplifying ‘high’ scores (e.g., reflecting the 6-7 score point range on the instructional quality dimensions of the CLASS, or the 3 score point range on the MQI) may sharpen raters’ accuracy. On the other hand, consistent with others’ conclusions (Guerrero-Rosada et al., 2021; McDoniel et al., 2022), the constructs that classroom observation tools are measuring may also benefit from ongoing experimentation and refinement, such as to consider greater focus on dimensions of curriculum, pacing, and coherence that support students’ opportunity to learn, but on the current versions of the instruments receive less emphasis.

More broadly, we argue that testing classroom observation protocols' ability to distinguish 'discriminant groups,' or groups of lessons whose instructional quality the protocols should in theory be able to distinguish, can provide useful evidence for ongoing validation work for classroom observation instruments. A validity argument approach implies that classroom observation scores should align with experts' views of lessons' instructional quality (e.g., Hill et al, 2011). Charalambous and Praetorius (2018) put forth the following set of propositions to aid in steering validation work for uses of classroom observation protocols:

(a) the work of teaching can be decomposed into a (comprehensive) set of quality characteristics; (b) these characteristics can be measured via observations; (c) appropriate scales and anchor points can be developed to capture the quality of these characteristics; (d) these characteristics form distinguishable latent constructs; (e) instructional quality scores obtained from classroom observation instruments can be generalized across relevant conditions; (f) students whose teachers score highly on the instructional quality measures will exhibit higher learning gains compared to students whose teachers score less highly; (g) instructional quality scores are related to other teacher-related constructs assumed to contribute to the work of teaching; and (h) results from different classroom observation instruments aiming to capture instructional quality should be closely related. (p. 542)

Applying classroom observation protocols to discriminant groups can be particularly generative in examining proposition (h). To the extent that ratings using an observation protocol do not converge with experts' established views on the relative instructional quality of a 'known groups' set of lessons, researchers can productively explore why the results differed from expectations, whether due to rater error, construct underrepresentation on the protocol, or other factors. The current study provides an illustration of how such an approach could be applied. Building out a contemporary corpus of shared, widely accessible mathematics lesson videos would be a valuable future step to help the field construct more common frames of reference for conducting such work.

Strengths, Limitations, and Future Research

Like all research studies, the current study had both strengths and limitations, which point towards fruitful pathways for future research. First, the corpus of TIMSS public-domain videos is small and was collected prior to the wave of reform initiatives in the United States that included the dissemination of the Common Core State Standards (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010). However, the TIMSS public domain videos were specifically assembled to provide a common frame of reference for cross-national research on teaching (Stigler, n.d.); analyzing these videos aligns with this goal. As such, we posit that the present analysis comprises a form of existence proof for the notion that examining the performance of classroom observation scores using discriminant groups of lessons can yield useful insights. In the future, an

updated version of the TIMSS Video Study that disseminated a new and expanded set of international mathematics lesson videos would certainly be generative to the field.

Additionally, the PA lessons were selected by TIMSS researchers to be reflective of the most salient aspects of classroom mathematics instruction visible in the sample of TIMSS video lessons; if the TIMSS research team failed at this task, conclusions about instructional quality seen in the PA videos would not be generalizable to the full restricted TIMSS 1999 Video Study dataset. However, we do not have indication to believe this occurred. As noted above, members of the original TIMSS country research teams, including code developers and coders, who would have been familiar with the original study coding, made PA video selections using a systematic approach with the explicit goal of selecting PA videos that exemplified the kinds of instruction visible in the full dataset. Because the PA lessons were not selected based on topics, we cannot make direct comparisons among teachers teaching the same lesson content. However, in prior work researchers using the TIMSS dataset have documented that even when teaching similar content, U.S. lessons tended to emphasize definitions and procedures, while Japanese lessons emphasized understanding relationships and applying mathematical concepts (Smith, 2011; Stigler & Hiebert, 1999); thus, the patterns identified here are unlikely to be driven simply by lesson content.

The use of multiple languages is a challenge in multinational research on teaching generally. We could not and did not have access to raters who were both certified on the MQI and CLASS instruments and spoke all of the original TIMSS languages; as such, we are unable to gauge the potential influence of rater language fluency. The TIMSS lessons were professionally videotaped and comprehensively translated using a detailed procedure (Jacobs et al., 2007); raters viewed these videos and read the English transcriptions to complete scoring. Larger studies using certified raters fluent in the original lesson languages would be a useful replication step. Lastly, while observations were conducted using two influential and widely used classroom observation instruments, we could not test every rubric. Future replication studies applying additional classroom observation tools to larger multi-national video pools would be useful to confirm and extend the current findings.

In sum, we used two contemporary classroom observation protocols to rate classroom mathematics lessons systematically chosen to be illustrative of the central facets of teaching in the country samples from Japan and the United States in the TIMSS 1999 Video Study, an important study in mathematics education research. Through qualitative examples unpacking how classroom observers rated segments of mathematics instruction from the United States and Japan, we illustrated how applying classroom observation protocols to ‘contrasted groups’ of lessons can surface areas where score patterns may differ from expectations, and thus point toward places for potential productive refinement to scoring tools and processes. Overall, the observed findings broaden the body of literature suggesting future research of value to refining classroom observation processes. Ultimately, continuously improving classroom observations can help policymakers and practitioners to more effectively evaluate the impacts of mathematics

education research interventions, and to direct teacher professional learning investments where they are needed most, contributing to strengthened long-run learning opportunities for children in mathematics.

Acknowledgements

I thank Cynthia Pollard for research assistance, and Charalambos Charalambous, Heather Hill, Corinne Herlihy, Jon Star, and Dan Koretz for helpful discussions and comments on earlier versions. All mistakes are mine.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Ayalon, M., & Rubel, L. H. (2022). Selecting and sequencing for a whole-class discussion: Teachers' considerations. *The Journal of Mathematical Behavior*, 66, 100958.
- Bell, C. A., Dobbelaer, M. J., Klette, K., & Visscher, A. (2019). Qualities of classroom observation systems. *School Effectiveness and School Improvement*, 30(1), 3-29.
- Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, 17(2-3), 62-87. doi: 10.1080/10627197.2012.715014
- Bell, C. A., Qi, Y., Croft, A. J., Leusner, D., McCaffrey, D. F., Gitomer, D. H., & Pianta, R. C. (2015). Improving observational score quality: Challenges in observer thinking. *Designing teacher evaluation systems: New guidance from the measures of effective teaching project*, 50-97.
- Berk, R. A. (1976). Determination of optional cutting scores in criterion-referenced measurement. *The Journal of Experimental Educational*, 4-9.
- Bostic, J., Lesseig, K., Sherman, M., & Boston, M. (2021). Classroom observation and mathematics education research. *Journal of Mathematics Teacher Education*, 24, 5-31.
- Brophy, J. (1986). Teacher influences on student achievement. *American Psychologist*, 41(10), 1069.
- Centers for Medicare and Medicaid Services Measures Management System. (2023). Measure evaluation criteria: validity. <https://mmshub.cms.gov/measure-lifecycle/measure-testing/evaluation-criteria/scientific-acceptability/validity>
- Charalambous, C. Y., & Praetorius, A. K. (2018). Studying mathematics instruction through different lenses: Setting the ground for understanding instructional quality more comprehensively. *ZDM*, 50, 355-366.
- Charalambous, C. Y., & Praetorius, A. K. (2022). Synthesizing collaborative reflections on classroom observation frameworks and reflecting on the necessity of synthesized frameworks. *Studies in Educational Evaluation*, 75, 101202.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281.
- Delaney, S. (2012). A validation study of the use of mathematical knowledge for teaching measures in Ireland. *ZDM*, 44, 427-441.
- Fernandez, C. & Yoshida, M. (2004). *Lesson study: A Japanese approach to improving mathematics teaching and learning*. Mahwah, NJ: Erlbaum.
- Geist, E. A. (2000). Lessons from the TIMSS videotape study. *Teaching Children Mathematics*, 7(3), 180-185.
- Guerrero-Rosada, P., Weiland, C., McCormick, M., Hsueh, J., Sachs, J., Snow, C., & Maier, M. (2021). Null relations between CLASS scores and gains in children’s language, math, and executive function skills: A replication and extension study. *Early Childhood Research Quarterly*, 54, 1-12.
- Hiebert, J., Gallimore, R., Garnier, H., Givvin, K. B., Hollingsworth, H., Jacobs, J., . . . Stigler, J. W. (2003). *Teaching mathematics in seven countries: Results from the TIMSS 1999 Video Study* (NCES Publication No. 2003-013). Washington, DC: U.S. Department of Education, National Center for Education Statistics.
- Hiebert, J., Gallimore, R., & Stigler, J. W. (2003). The new heroes of teaching. *Education Week*, 23(10), 42-56.
- Hiebert, J., Stigler, J. W., Jacobs, J. K., Givvin, K. B., Garnier, H., Smith, M., . . . Gallimore, R. (2005). Mathematics teaching in the United States today (and tomorrow): Results from the TIMSS 1999 video study. *Educational Evaluation and Policy Analysis*, 27(2), 111-132.
- Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, 26(4), 430-511.
- Hill, H. C., Charalambous, C. Y., Blazar, D., McGinn, D., Kraft, M. A., Beisiegel, M., ... & Lynch, K. (2012). Validating arguments for observational instruments: Attending to multiple sources of variation. *Educational Assessment*, 17(2-3), 88-106.
- Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, 48(3), 794-831.
- Hofkens, T., Pianta, R. C., & Hamre, B. (2023). Teacher-student interactions: theory, measurement, and evidence for universal properties that support students’ learning across countries and cultures. In *Effective Teaching Around the World: Theoretical, Empirical, Methodological and Practical Insights* (pp. 399-422). Cham: Springer International Publishing.
- Jacobs, J. K., Hollingsworth, H., & Givvin, K. B. (2007). Video-based research made “easy”: Methodological lessons learned from the TIMSS video studies. *Field Methods*, 19(3), 284-299.
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological bulletin*, 112(3), 527.
- Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching*. <http://www.metproject.org/reports.php>
- Klette, K., & Blikstad-Balas, M. (2018). Observation manuals as lenses to classroom teaching: Pitfalls and possibilities. *European Educational Research Journal*, 17(1), 129-146.
- Lampert, M. (2001). *Teaching problems and the problems of teaching*. New Haven: Yale University Press.
- La Paro, K. M., Pianta, R. C., & Stuhlman, M. (2004). The classroom assessment scoring system: Findings from the

- prekindergarten year. *The Elementary School Journal*, 104(5), 409-426.
- LessonLab. (n.d.-a). JP3 Lesson Graph.
https://static1.squarespace.com/static/59df81ea18b27ddf3bb4abb5/t/59fcd8769140b790c66b6422/1509742710519/JP3+Lesson+Graph_0.pdf
- LessonLab. (n.d.-b). JP3 Lesson Plan.
https://static1.squarespace.com/static/59df81ea18b27ddf3bb4abb5/t/59fcd8940d9297c985fda01f/1509742740754/JP3+Lesson+Plan_0.pdf
- LessonLab. (n.d.-c). JP3 Teacher Comments.
<https://static1.squarespace.com/static/59df81ea18b27ddf3bb4abb5/t/5ca54f9015fcc0468d2c2739/1554337680488/JP3+Teacher+Comments.pdf>
- LessonLab. (n.d.-d). JP3 Transcript.
<https://www.timssvideo.com/jp3-solving-inequalities>
- LessonLab. (n.d.-e). US2 Lesson Graph.
<https://static1.squarespace.com/static/59df81ea18b27ddf3bb4abb5/t/59fd08836c3194747a386a9f/1509755012151/US2+Lesson+Graph.pdf>
- LessonLab. (n.d.-f). US2 Teachers Comments.
<https://static1.squarespace.com/static/59df81ea18b27ddf3bb4abb5/t/5bc512b6104c7b27edd10d23/1539642038921/US2+Teacher+Comments.pdf>
- LessonLab. (n.d.-g). US2 Transcript. <https://www.timssvideo.com/us2-writing-variable-expressions#tabs-2>
- Leung, F. K. S. (2005). Some characteristics of East Asian mathematics classrooms based on data from the TIMSS 1999 video study. *Educational Studies in Mathematics*, 60, 199-215.
- Leyva, D., Weiland, C., Barata, M., Yoshikawa, H., Snow, C., Treviño, E., & Rolla, A. (2015). Teacher-child interactions in Chile and their associations with prekindergarten outcomes. *Child Development*, 86(3), 781-799.
- Mackie, D.M., & Smith, E. (1998). Intergroup relations: insights from a theoretically integrative approach. *Psychological Review*, 105(4): 499-529.
- Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., ... & Howes, C. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development*, 79(3), 732-749.
- McDoniel, M. E., Townley-Flores, C., Sulik, M. J., & Obradović, J. (2022). Widely used measures of classroom quality are largely unrelated to preschool skill development. *Early Childhood Research Quarterly*, 59, 243-253.
- National Council of Teachers of Mathematics (NCTM). (2014). *Principles to actions: Ensuring mathematical success for all*. Reston, VA: Author.
- National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for Mathematics*. Washington, DC: Authors.

- Nittler, K. (2020). How evaluation ratings impact teacher pay. <https://www.nctq.org/blog/How-evaluation-ratings-impact-teacher-pay>
- Office of Head Start. (2015). *Use of CLASS in Head Start*. <http://eclkc.ohs.acf.hhs.gov/hslc/hs/sr/class>
- Pakarinen, E., Lerkkanen, M. K., Poikkeus, A. M., Kiuru, N., Siekkinen, M., Rasku-Puttonen, H., & Nurmi, J. E. (2010). A validation of the classroom assessment scoring system in Finnish kindergartens. *Early Education and development, 21*(1), 95-124.
- Perlman, M., Falenchuk, O., Fletcher, B., McMullen, E., Beyene, J., & Shah, P. S. (2016). A systematic review and meta-analysis of a measure of staff/child interaction quality (the classroom assessment scoring system) in early childhood education and care settings and child outcomes. *PloS One, 11*(12), e0167660.
- Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*(2), 109-119.
- Pianta, R.C., Hamre, B., & Mintz, S. (2012). *Upper elementary and secondary CLASS technical manual*.
- Praetorius, A. K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of three basic dimensions. *ZDM, 50*, 407-426.
- Praetorius, A. K., Klieme, E., Kleickmann, T., Brunner, E., Lindmeier, A., Taut, S., & Charalambous, C. (2020). *Towards developing a theory of generic teaching quality. Origin, current status, and necessary next steps regarding the Three Basic Dimensions Model* (pp. 15-36).
- Richland, L. E. (2015). Linking gestures: Cross-cultural variation during instructional analogies. *Cognition and Instruction, 33*(4), 295-321.
- Santagata, R., Kersting, N., Givvin, K. B., & Stigler, J. W. (2010). Problem implementation as a lever for change: An experimental study of the effects of a professional development program on students' mathematics learning. *Journal of Research on Educational Effectiveness, 4*(1), 1-24.
- Schoen, H. L., Cebulla, K. J., Finn, K. F., & Fi, C. (2003). Teacher variables that relate to student achievement when using a standards-based curriculum. *Journal for Research in Mathematics Education, 34*(3), 228-259.
- Schoenfeld, A. S. (1998). Toward a theory of teaching-in-context. *Issues in Education, 4*(1), 1-95.
- Smith, M. (2011). A procedural focus and a relationship focus to algebra: How US teachers and Japanese teachers treat systems of equations. In J. Cai & E. Knuth (Eds.), *Early algebraization: A global dialogue from multiple perspectives* (pp. 511-528). Springer Berlin Heidelberg.
- Smith, M., & Stein, M. K. (2018). *5 Practices for orchestrating productive mathematics discussion*. National Council of Teachers of Mathematics.
- Star, J. R., Pollack, C., Durkin, K., Rittle-Johnson, B., Lynch, K., Newton, K., & Gogolen, C. (2015). Learning from comparison in algebra. *Contemporary Educational Psychology, 40*, 41-54.
- Stein, M. K., Engle, R. A., Smith, M. S., & Hughes, E. K. (2008). Orchestrating productive mathematical discussions: Helping teachers learn to better incorporate student thinking. *Mathematical Thinking and Learning, 10*, 313-340.
- 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.
- Stigler, J. W. (n.d.). Collecting the public use lessons. <http://www.timssvideo.com/>
- Stigler, J. W., & Hiebert, J. (1999). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom*. New York: Simon & Schuster.
- Tekkumru-Kisa, M., Stein, M. K., & Doyle, W. (2020). Theory and research on tasks revisited: Task as a context for students' thinking in the era of ambitious reforms in mathematics and science. *Educational Researcher*, 49(8), 606-617.
- Verleghe, P. W. (2007). Home country bias in product evaluation: the complementary roles of economic and socio-psychological motives. *Journal of International Business Studies*, 38, 361-373.
- Walsh, K., & Ross, E. (2019). *NCTQ State of the states 2019: Teacher and principal evaluation policy*. <https://www.nctq.org/publications/State-of-the-States-2019:-Teacher-and-Principal-Evaluation-Policy>
- White, M. C. (2018). Rater performance standards for classroom observation instruments. *Educational Researcher*, 47(8), 492-501.
- White, M., & Klette, K. (2024). Signal, error, or bias? exploring the uses of scores from observation systems. *Educational Assessment, Evaluation and Accountability*, 1-24.

Appendix

Table S1

Descriptive Results of PA Lessons' Classroom Observation Scores

	Instructional Outcomes	
	Japan Mean (SD)	United States Mean (SD)
CLASS Instructional Support Dimensions		
Content Understanding	4.25 (0.29)	4.56 (0.65)
Analysis and Problem-Solving	4.17 (0.68)	3.79 (0.97)
Quality of Feedback	3.70 (1.10)	4.63 (0.96)
Instructional Dialogue	3.75 (0.97)	4.61 (0.55)

Note: Means reflect averages of country mean scores on each of the CLASS Instructional Support dimensions. CLASS codes are on a scale of 1-7 (low-high).

Corresponding Author Contact Information:

Author name: Kathleen Lynch

Department: Educational Psychology

University, Country: University of Connecticut, USA

Email: kathleen_lynch@uconn.edu

Please Cite: Lynch, K. (2024). Seeing through the eyes of classroom observers: The case of rating contrasted groups of lessons with classroom observation measures. *Journal of Research in Science, Mathematics and Technology Education*, 7(2), 47-77. DOI: <https://doi.org/10.31756/jrsmte.723>

Copyright: This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Conflict of Interest: None reported.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Data Availability Statement: The public-use TIMSS Video Study data are freely available online at <http://www.timssvideo.com/>.

Ethics Statement: This study did not use human or animal subjects.

Author Contributions: All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Received: February 27, 2024 ▪ Accepted: May 10, 2024