

Chapter 16

Teacher Effectiveness in Multiple Lenses: Secondary Analysis Lessons in the Measures of Effective Teaching Project



James Ko , Zhijun Chen, Jieyan Celia Lei, and Ridwan Maulana 

Abstract Improving teaching quality to enhance learning has become critical for academics, practitioners, and policymakers. However, very few studies compared the same lessons with various classroom observation instruments to examine whether classroom characteristics in different instruments are similar.

This project aimed to conduct a secondary lesson observation analysis on 423 lesson videos selected from 14,000+ lesson videos previously collected in the *Measures of Effective Teaching* (MET) project. The analysis provided new data on the same lessons previously studied by the MET researchers but observed with two instruments. One internationally validated instrument was used in the international project by Maulana et al. (Sch Eff Sch Improv, 1–32, 2021) to explore the generic teaching characteristics in different countries, while the other instrument was developed purposively to characterise the differences between effective and inspiring teaching.

J. Ko (✉)

Department of Education Policy and Leadership, The Education University of Hong Kong,
Tai Po, Hong Kong
e-mail: jamesko@eduhk.hk

Z. Chen

Department of Education, University of Bath, Bath, UK

J. C. Lei

Department of Education, Shaoyang University, Shaoyang, China

Department of Education Policy and Leadership, Education University of Hong Kong,
Tai Po, Hong Kong

R. Maulana

Department of Teacher Education, University of Groningen, Groningen, The Netherlands
e-mail: r.maulana@rug.nl

© The Author(s) 2023

R. Maulana et al. (eds.), *Effective Teaching Around the World*,
https://doi.org/10.1007/978-3-031-31678-4_16

339

The results allowed data comparisons across different projects that are not readily comparable because they used various classroom observation instruments. The results informed the relationships between effective and inspiring teaching.

Keywords Teacher effectiveness · Teaching quality · Classroom instrument · Video lesson analysis

1 Introduction

Beliefs about what constitutes ‘good’ or ‘high’ quality practice in teaching can vary markedly for different age groups of students, at other times and in different contexts. ‘Effectiveness’ is a contested term that can evoke strong emotions because perceived effectiveness links with notions of professional competency and high-stakes accountability in some countries. Researchers may question individual teachers’ beliefs about their professional autonomy. Notions of what constitutes high quality or good teaching, or the idea that teaching is an art or a craft rather than a science, are sometimes used to raise concerns with narrower concepts of effectiveness.

Researchers recognise the importance of effective teaching behaviour for student outcomes, but most teachers still struggle to implement complex teaching skills in their daily classroom practices. The Measures of Effective Teaching (MET) project (Bill & Melinda Gates Foundation, 2018) has provided the research communities with the most extensive dataset on classroom observation with an easily accessible video library for secondary data analysis. To our knowledge, this is the first study to connect two large-scale classroom observation studies through secondary data analysis of selected lesson videos with the same instruments.

2 Theoretical Framework

2.1 *Examining Teacher Effectiveness Through Classroom Observations*

Teacher effectiveness research is a branch of educational effectiveness research, focusing mainly on variations in teaching quality on student outcomes. Value-added measures, classroom observations, and student surveys are familiar sources of information and data about teachers’ behaviour and classroom practices that can be drawn upon to provide evidence to inform our understanding of teacher effectiveness (Bacher-Hicks et al., 2019). If student outcomes are the essential criteria for teacher effectiveness, the question remains about what kinds of outcomes,

objectives, and goals can be achieved by teachers and schools. We clearly cannot go on endlessly adding more objectives and more content for teachers and schools and still expect them to succeed. Teacher effectiveness beyond the classroom level (e.g., Cheng, 1996; Cheng & Tsui, 1998) is not practically appealing to practitioners because such a conceptualisation could obscure the focus of the teacher's role and duties in teaching. In practice, teachers and schools often prefer to restrict teacher evaluation to specific objectives in teaching.

For evaluating teaching quality, while gains in cognitive and non-cognitive domains of student achievement are tentative, a plea for meeting cognitive purposes and obtaining higher academic attainments as the criteria for the effectiveness of education in schools often sounds more appealing. Bacher-Hicks et al. (2019) argued for value-added measures as unbiased predictors of teacher performance in experimental conditions where students were assigned randomly to different classrooms. However, value-added measures are not unbiased as assumed because they tend to shift when different tests assess student achievements (Grossman et al., 2014).

While a classroom observation approach cannot adjust for classroom composition, it has two obvious advantages apart from easily accessible applications in naturally occurring settings. First, it allows ready comparisons across grades and subjects without relying on reliable, standardised tests. Second, a classroom observation approach looks at teacher effectiveness from a different angle by allowing the observers or evaluators to associate the observed behaviours with various aspects of the student learning process, such as student engagement in class and students' self-reported behaviours or learning characteristics (e.g., Clunies-Ross et al., 2008; Helmke et al., 1986; Virtanen et al., 2015).

2.2 *Comparisons of Classroom Observation Instruments*

Classroom observation is a powerful method to collect data on teacher behaviours in class. Numerous sources of information and data about teachers' behaviour and classroom practices can be drawn upon to provide evidence to inform our understanding of teacher effectiveness. A standard method in a classroom observation approach to teacher effectiveness research is to observe different teachers' teaching practices by independent observers. For example, this paper compared the results obtained from different high-inference instruments to capture aspects of teaching dimensions hypothesised to be operated at the classroom level.

We can compare different classroom observation instruments of similar nature (i.e., for generic teaching behaviours) for different lessons in a single study (Day et al., 2008; Kington et al., 2014), different classroom observation instruments of similar nature for the same lessons in a single study (e.g., Ko, 2010; Ko et al., 2015; Lei et al., 2023; Sammons et al., 2014, 2016), and different classroom observation instruments of different nature (i.e., effective vs inspiring teaching behaviours) for different lessons in a single study (Ko et al., 2019a, b, 2016; Zhao & Ko, 2022).

However, the measurement strategy of teacher effectiveness in the MET project was unique as it compared different classroom instruments that differed in specificity. It involved comparisons of generic teacher behaviours by the *Framework for Teaching* (Danielson, 2013) and the *Classroom Assessment Scoring System* (CLASS) (Pianta et al., 2012) and of subject-specific ones by the *Mathematical Quality of Instruction* (MQI) (Hill et al., 2008; The Learning Mathematics for Teaching Project, 2011), the *Protocol for Language Arts Teaching Observations* (PLATO), the *Quality of Science Teaching* (QST) (Schultz & Pecheone, 2014), and the *UTeach Teacher Observation Protocol* (UTOP) (Walkington & Marder, 2014, 2018).

The challenges of developing and comparing quality teacher observation systems lie in establishing rater reliability and making the instruments more generalisable across contexts (Hill et al., 2012; Liu et al., 2019). For example, despite its wide application, the CLASS was not adequately validated without revisions in Hafen et al. (2015), where the lesson videos were collected in various projects. Wallace et al. (2020) also reported that the CLASS failed to discriminate classroom management quality, with most teachers' scores clustering around the most positive ranges of effectiveness. The present study differed from the heuristic comparison of classroom observation instruments by Bell et al. (2019) in that we observed the same lessons with different instruments. Bell et al.'s (2019) comparison was crude and non-quantitative, as all instruments they compared shared ten similar teaching dimensions.

Secondary data analysis on the MET data should provide quantitative evidence for instrument comparisons, but to date, we still cannot find any study exploring this. We intended to fill this gap with this study and were motivated to conduct secondary data analysis with the same classroom observation instrument in the international collaborative ICALT3 project (Maulana et al., 2021) so that the new data on the selected sample would form a part of the enlarged study to inform the measurement invariance of teaching quality (Krammer et al., 2020; Maulana et al., 2021).

2.3 Video Lesson Analysis¹

The TIMSS 1995 video study by Stigler et al. (1999) was a pioneer and exemplar in using video data to explore teaching characteristics and patterns cross-culture beyond qualitative coding to provide quantitative analysis for hypothesis testing

¹We deliberately left out discussing the OECD TALIS video study (McCann et al., 2020; OECD, 2017) for this section because the study was still on-going by the time of writing. Although a very elaborated observation instrument was developed for the study, it has not been applied to any other study beyond the OECD. We also excluded TEACH (World Bank, 2019) for its limited application in research to date, but a chapter by (Lei et al., 2023, this volume) that compares TEACH and ICALT can be found in this volume.

(Jacob et al., 1999; Stigler et al., 2000). The initial sample included 231 mathematics lessons from Germany, Japan, and the United States, selected from a nationally representative sample of eighth-grade students and classrooms participating in the 1994–95 TIMSS assessments. The TIMSS 1999 video study expanded to include Australia, the Czech Republic, Hong Kong SAR, the Netherlands, Switzerland, and the United States (Hiebert et al., 2003). In the video study on science teaching, the participating countries included Australia, the Czech Republic, Japan, the Netherlands, and the United States (Roth et al., 2006). However, no observation instruments were developed or adopted for observations in these studies. Only a portion of the lesson videos are publicly available for secondary data analysis, so our purpose should provide new data for the ICALT3 with the MET data.

Apart from the MET project, only a few studies in the literature used lesson videos to conduct lesson observation to inform teaching practice and performance (e.g., Hafen et al., 2015; Ko et al., 2015, 2016). Secondary data analysis makes instrument comparisons feasible if a lesson is videotaped for observation, as in the MET project, providing opportunities to observe the same lessons again at different times and research contexts.

3 Methods

3.1 Data Collection

The current study used both the original data of the CLASS and new observational data using two classroom observations to compare classroom characteristics.

3.2 Raters

The second and third authors conducted the majority of the lessons. The third author assisted the second author as a research assistant to use ICALT in another project (Lei et al., 2023). When these research assistants shared the secondary video analysis, they passed the training session and conducted two calibrations. One English for Language Arts (ELA) lesson and one Math lesson were observed and scored twice by each rater in each calibration. After each calibration, they conducted an interrater reliability test and proceeded with the lesson observation when Krippendorff's alpha (2004) increased from .52 to .73 for the ELA lesson and from .55 to .82 for the ELA Math lesson.

3.3 Video Samples

3.3.1 Original Lesson Videos of the MET Study

The *Measurement of Effective Teaching* (MET) project was a large project funded by the Bill & Melinda Gates Foundation (2018). Around 2700 teachers from 10 districts in the United States teaching science, English, and math across 4–9 grades participated from 2009–2010 and 2010–2011 (Bill & Melinda Gates Foundation, 2018). Each teacher was videotaped during the lessons one to four times over a year. After training, the lessons were divided into segments and coded in 20-min segments by their administrator and peer observers using different classroom observation instruments. Despite its scale, teachers, classrooms, schools and districts in the MET project were not randomised.

4 Current Secondary Data Analysis

Among these different instruments, the *CLASS* was used for all lessons in the MET project and the most studied instrument outside the U.S.A. (e.g., Taut et al., 2019 in Chile; Pöysä et al., 2019 in Finland; Havik & Westergård, 2020 in Norway). The *CLASS* was assumed to be a reliable reference for selecting lesson videos for secondary data analysis with two new classroom observation instruments, the *International Comparative Analysis of Learning and Teaching* (ICALT) and the *Comparative Analysis of Effective Teaching and Inspiring Teaching* (CETIT). Thus, in this study, we selected four hundred twenty-three lessons proportional to the staine distribution (Clark-Carter, 2005) of the percentiles of the aggregated mean scores of the various teaching dimensions of the *CLASS*.² We also limited the sample to secondary school lessons (i.e., 7–9) and English and mathematics only. Two lessons were excluded due to low video quality. Three trained raters observed nine lessons for calibrations first and started secondary observations after inter-rater reliability was over 90%. Each observer was assigned randomly to observe different lessons. The total numbers of segment, video, rater, and teacher are summarised in Table 16.1.

²The distribution of *CLASS* scores in the MET project was normal. Thus, using staines to select sample lessons for the secondary analysis retained a similar normal distribution.

Table 16.1 The total numbers of segment, video, rater and teacher of the original data in the MET project and 423 chosen in this project

	Original MET	Current secondary data analysis
Segment	3	1–3
Lesson	14,179	423
Rater	419	3
Teacher	1594 (4 lessons)	217 (1 lesson); 103 (2 lessons)
Year	1st (2010); 2nd (6294)	1st (318); 2nd (105)
Grade	4th–9th	7th–9th
Subject	2	2

4.1 Instruments

4.1.1 CLASS

The *CLASS* in the MET project has an additional dimension, *Instruction Dialogue*, in addition to its original version with ten³ dimensions of teaching quality: *Positive Climate*, *Negative Climate*, *Teacher Sensitivity*, *Adolescent Perspectives*, *Behaviour Management*, *Productivity*, *Instructional Learning Formats*, *Content Understanding*, *Analysis and Inquiry*, *Quality of Feedback*, and one dimension of *Learner Engagement* (Pianta & Hamre, 2009). Each lesson was divided into one, two, or three segments, each rated independently by a different rater on a 7-point Likert scale representing low to high levels.

4.1.2 ICALT

Originally developed as an instrument for inspection to capture generic teaching behaviours (van de Grift, 2007, 2014), the *ICALT* has expanded into thirty-two high-inference teaching indicators categorised into six domains: *Safe and stimulating learning climate*, *Efficient organisation*, *Clear and structured instructions*, *Intensive and activating teaching*, *Adjusting instructions and learner processing to inter-learner differences*, and *Teaching learning strategies*. The *ICALT* also contained a three-item (e.g., ‘...take an active approach to learn’) student learning domain to document learner engagement during classroom observations. Three observers completed classroom observation for each lesson and rated the items based on teachers’ performance on a 4-point scale, from ‘mostly weak’ to ‘mostly strong.’

³In Pianta and Hamre (2009), there was a dimension of *Procedures and skills*, which was not in the MET.

4.1.3 CETIT

Based on the teaching aspects characterised as inspiring teaching by Sammons et al. (2014), Ko et al. (2016) used the Delphi method to finalise and validate the *CETIT*. This new high-inference classroom observation instrument consisted of sixty-eight descriptive statements that included effective and inspiring teaching domains. According to Ko et al. (2016), inspiring teaching includes four aspects: *Flexibility*, *Teaching reflective thinking*, *Innovative teaching*, and *Teaching collaborative learning*. Teaching behaviours corresponding to these inspiring teaching domains include “*The teacher allowed options for students in their seatwork*,” “... *asked students to comment on his/her viewpoint*,” “... *used ICT in teaching*,” and “... *told students how to share their work in a task*.” While *Teaching reflective thinking* and *Teaching collaborative learning* were two distinctive classroom practices in the *CETIT*, they were conceived as a single characteristic by Sammons et al. (2014). *Dimensions Assessment for learning* and *Professional Knowledge and expectations* are two unique teaching aspects in the *CETIT* (i.e., not found in the *CLASS* or the *ICALT*). They were found to cluster with other teaching domains of effective teaching (Ko et al., 2016, 2019a, b). For this study, two new dimensions, *Engagement in exploratory learning* and *Engagement in knowledge consolidation*, developed by Piburn and Sawada (2000), were adopted to test whether the learner dimensions in different instruments might favour the teaching dimensions of the classroom observation instruments to which they belong.

4.2 Data Analysis

For all three instruments, the means, standard deviations, and reliability tests were conducted in SPSS 20. Confirmatory factor analyses (CFA) were conducted in MPlus 7. The original three-factor model of the *CLASS* was tested first, followed by one-factor and two-factor models for comparison. For the *ICALT*, a six-factor model was tested with the theoretical structure. Three CFA models were tested on the *CETIT*: (a) an eight-factor model on effective teaching, (b) a four-factor model on inspiring teaching, and (c) a 12-factor full model. Multiple good fit indices were selected as the criteria suggested by Tabachnick et al. (2007) for evaluating the CFA models: (a) the Root mean square error of approximation (RMSEA) below .08, (b) a Comparative fit index (CFI) above .95, (c) standardised root mean square residual under .08, and (d) χ^2/df to be under 2.

5 Findings

5.1 Descriptive Statistics

5.1.1 CLASS

The overall results shown in Table 16.2 are consistent with the CLASS results in the literature. *Instructional support* was the weakest domain. At the dimension level, the average scores were relatively low for *Negative Climate* ($M = 1.47$, $SD = .63$) and *Analysis and Inquiry* ($M = 2.42$, $SD = .90$). In contrast, Dimensions *Behavior Management* ($M = 5.72$, $SD = .98$) and *Productivity* ($M = 5.54$, $SD = .93$) were scored relatively higher than all other dimensions. Table 16.2 indicated that the reliability for each domain, *Emotional Support*, *Classroom Organisation*, or *Instructional Support*, was acceptable as a subscale and the full-scale CLASS ($\alpha > .7$). There were no reliability scores for dimensions because they were single indicators.

5.1.2 ICALT

For ICALT, the means of *Adjusting Instructions and Learner Processing to Inter-Learner Differences* ($M = 1.68$, $SD = .42$) and *Teaching Learning Strategies* ($M = 1.45$, $SD = .40$) were low because they were rare in the sampled lessons. The reliability test results indicated a high level of internal consistency for the full scale of ICALT ($\alpha = .87$). Still, as depicted in Table 16.3, half of the ICALT domains have

Table 16.2 Means, standard deviations and Cronbach's Alphas (α) of CLASS

Dimension/Domain	Mean	SD	Cronbach
Positive climate	4.14	.97	NA
Negative climate	1.47	.63	NA
Teacher sensitivity	3.87	.97	NA
Regard for adolescent perspectives	2.93	1.03	NA
Behaviour management	5.72	.98	NA
Productivity	5.54	.93	NA
Instructional learning formats	3.85	.91	NA
Content understanding	3.59	.98	NA
Analysis and inquiry	2.42	.90	NA
Quality of feedback	3.21	1.02	NA
Instructional dialogue	2.94	1.06	NA
Student engagement	4.50	.92	NA
<i>Emotional support</i>	4.12	.70	.76
<i>Classroom organization</i>	5.04	.75	.72
<i>Instructional support</i>	3.04	.87	.90
Full scale	4.02	.68	.90

Table 16.3 Mean, standard deviation and Cronbach's Alpha (α) of ICALT

Dimension	Mean	SD	Cronbach
Safe and stimulating learning climate	3.04	.45	<u>.58</u>
Efficient organisation	3.30	.40	<u>.65</u>
Clear and structured instructions	2.88	.41	.76
Intensive and activating teaching	2.29	.41	<u>.56</u>
Adjusting instructions and learner processing to inter-learner differences	1.68	.42	<u>.49</u>
Teaching learning strategies	1.45	.40	.70
Learner engagement	2.93	.49	.79
Full scale	2.40	.30	.87

Table 16.4 Mean, standard deviation and Cronbach's Alpha (α) of CETIT

Dimension	Mean	SD	Cronbach
Enthusiasm for teaching	3.17	.49	.84
Positive relationships with students	2.72	.47	.72
Purposeful and relevant teaching	1.89	.42	<u>.57</u>
Safe classroom climate	2.86	.46	<u>.58</u>
Stimulating learning environment	1.43	.39	<u>.23</u>
Positive classroom management	3.37	.45	.73
Assessment for learning	2.18	.35	<u>.61</u>
Professional knowledge and expectations	2.88	.40	<u>.60</u>
Flexibility	1.13	.27	<u>.58</u>
Teaching reflective thinking	1.35	.34	.70
Teaching collaborative learning	1.43	.54	<u>.68</u>
Innovative teaching	1.32	.35	<u>.59</u>
Engagement in exploratory learning	1.68	.31	<u>.54</u>
Engagement in knowledge consolidation	2.33	.42	<u>.49</u>
Full scale	2.10	.26	.93

reliability below .7, the threshold acceptable in education research (Taber, 2018): *Safe and Stimulating Learning Climate* ($\alpha = .58$), *Intensive and Activating Teaching* ($\alpha = .56$) and *Adjusting Instructions and Learner Processing to Inter-Learner Differences* ($\alpha = .49$).

5.1.3 CETIT

The result suggested that the full scale with all 68 items was highly consistent ($\alpha = .93$). The result also indicated good reliabilities in most of the CETIT dimensions. Besides, there was an unacceptable internal consistency of the subscale *Stimulating Learning Environment* ($\alpha = .23$), with a relatively lower score average ($M = 1.43$, $SD = .39$). The four subscales with reliability close to the .6-threshold included *Flexibility* ($\alpha = .58$), *Purposeful and Relevant Teaching* ($\alpha = .57$), *Safe Classroom Climate* ($\alpha = .58$), and *Innovative Teaching* ($\alpha = .59$) (Table 16.4).

6 Correlations of Factors of Three Instruments

Table 16.5 displays the Pearson correlation coefficients of the teacher dimensions, learner engagement dimensions, and the whole scale. As correlations are sensitive to sample size, we should focus on the association's magnitude or strength. In general, a coefficient between .4 and .6 indicates a moderate strength. While a value above .6 suggests a strong association, a value between .2 and .4 is weak to mild. Values below .2 are considered weak even though the correlation may be statistically significant.

Most teaching dimensions of the *CLASS* were correlated significantly only with other dimensions of the same scale, but teaching dimensions of *ICALT* and *CETIT* correlated with other dimensions of each other scale. All eleven *CLASS* dimensions suggested weak or no correlations with the *ICALT* and *CETIT* dimensions. In the *ICALT*, the result indicated that the domain *Teaching Learning Strategies* did not correlate with three domains in the *ICALT*: *Safe and Stimulating Learning Climate*, *Efficient Organisation*, and *Clear and Structured Instructions*.

In the *CETIT*, the dimension *Innovative Teaching* showed no correlation with the other nine dimensions, except for *Flexibility* ($r = .271, p < .01$) and *Teaching Reflective Thinking* ($r = .280, p < .01$). All three domains were classified as inspiring teaching practices. In contrast, other *CETIT* dimensions were correlated significantly with most *ICALT* dimensions. Comparing the subscales of student engagement in the *CLASS*, *ICALT*, and *CETIT*, *Learner Engagement* in the *ICALT* showed stronger correlations with more teaching dimensions, six in the *CLASS*, five in the *ICALT*, and ten in the *CETIT*. *Learner Engagement* in the *ICALT* was also weakly associated with *Student Engagement* in the *CLASS* ($r = .239, p < .01$), *Engagement in Exploratory Learning* ($r = .481, p < .01$) and *Engagement in Knowledge Consolidation* ($r = .608, p < .01$) in the *CETIT*.

6.1 Comparing Confirmatory Factor Models of Three Instruments

6.1.1 CLASS

Except for the original three-factor models of the *CLASS*, the one-factor and two-factor models were also built up to investigate a better factor structure of the *CLASS* based on the sampled lessons. In all three models, the two-factor model showed a relatively better model fit than the other one-factor model and the original three-factor model. The one-factor model of the *CLASS* suggested poor model fit to the data, $\chi^2(54) = 846.491, p < .001, CFI = .768, RMSEA = .186$, but interestingly, the theoretical three-factor model of the *CLASS* had the worst fit indices, $\chi^2(41) = 774.629, p < .001, CFI = .761, RMSEA = .205$. In contrast, the two-factor

Table 16.5 Correlations between CLASS, ICALT & CETIT

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34		
1 CLASS_Pedagogic Climate	1																																			
2 CLASS_Students' Climate	0.29**	1																																		
3 CLASS_Instructional Quality	0.40**	0.35**	1																																	
4 CLASS_Resources Management	0.60**	0.38**	0.68**	1																																
5 CLASS_Behavior Management	0.24**	0.65**	0.20**	0.11**	1																															
6 CLASS_Pedagogical	0.59**	0.19**	0.61**	0.69**	0.30**	0.39**	1																													
7 CLASS_Instructional Learning Formats	0.65**	0.17**	0.54**	0.57**	0.34**	0.72**	0.1	1																												
8 CLASS_Content Understanding	0.51**	0.11*	0.60**	0.62**	0.19**	0.25**	0.53**	0.47**	1																											
9 CLASS_Analysis and Inquiry	0.51**	0.12**	0.66**	0.59**	0.20**	0.21**	0.65**	0.75**	0.63**	1																										
10 CLASS_Quality of Feedback	0.57**	0.16**	0.59**	0.70**	0.21**	0.29**	0.43**	0.69**	0.70**	0.61**	1																									
11 CLASS_Instructional Dialogue	0.26**	0.02	0.75**	0.29**	0.04*	0.06	0.22**	0.17**	0.25**	0.21**	0.76**	1																								
12 ICALT_0-6th and remaining learning climate	0.13**	0.07*	0.17**	0.14**	0.30**	0.15**	0.15**	0.11**	0.24**	0.1	0.62**	0.1	1																							
13 ICALT_0-6th and remaining resources	0.19**	0.04*	0.22**	0.15**	0.19**	0.19**	0.20**	0.22**	0.22**	0.22**	0.35**	0.29**	0.29**	1																						
14 ICALT_0-6th and remaining pedagogy	0.19**	0.04*	0.22**	0.15**	0.19**	0.19**	0.20**	0.22**	0.22**	0.22**	0.35**	0.29**	0.29**	0.35**	1																					
15 ICALT_6th-9th and remaining pedagogy	0.19**	0.04*	0.22**	0.15**	0.19**	0.19**	0.20**	0.22**	0.22**	0.22**	0.35**	0.29**	0.29**	0.35**	0.66**	1																				
16 ICALT_6th-9th and remaining resources	0.19**	0.04*	0.22**	0.15**	0.19**	0.19**	0.20**	0.22**	0.22**	0.22**	0.35**	0.29**	0.29**	0.35**	0.66**	0.66**	1																			
17 ICALT_6th-9th and remaining pedagogy	0.19**	0.04*	0.22**	0.15**	0.19**	0.19**	0.20**	0.22**	0.22**	0.22**	0.35**	0.29**	0.29**	0.35**	0.66**	0.66**	0.66**	1																		
18 ICALT_6th-9th and remaining resources	0.19**	0.04*	0.22**	0.15**	0.19**	0.19**	0.20**	0.22**	0.22**	0.22**	0.35**	0.29**	0.29**	0.35**	0.66**	0.66**	0.66**	0.66**	1																	
19 PERE_Perceived relationships with students	0.03*	0.09*	0.14**	0.18**	0.017	0.04*	0.09*	0.17**	0.12**	0.09*	0.08*	0.08*	0.08*	0.08*	0.08*	0.08*	0.08*	0.08*	0.08*	1																
20 PERE_Professional self-efficacy	0.035	0.04	0.051	0.14**	-0.066	0.021	0.091	0.097	0.02	0.03	0.04*	-0.02	-0.001	0.012	0.11**	0.11**	0.12**	0.12**	0.12**	0.12**	1															
21 PERE_Preparedness of relevant teaching	0.12**	-0.04	0.14**	0.21**	0.074	0.13**	0.19**	0.23**	0.079	0.14**	0.21**	0.09**	0.17**	0.15**	0.43**	0.42**	0.42**	0.42**	0.42**	0.42**	0.42**	1														
22 PERE_Self-efficacy (classroom)	0.10**	0.014	0.13**	0.18**	0.016	0.061	0.11*	0.20**	0.12*	0.17**	0.13**	0.15**	0.13**	0.13**	0.25**	0.25**	0.25**	0.25**	0.25**	0.25**	0.25**	1														
23 PERE_Student learning environment	0.16**	0.19**	0.19**	0.13**	0.14**	0.092	0.17**	0.15**	0.17**	0.12*	0.12*	0.12*	0.12*	0.12*	0.20**	0.20**	0.20**	0.20**	0.20**	0.20**	0.20**	0.20**	1													
24 PERE_Pedagogical resources management	0.042	0.065	0.11*	0.091	-0.046	0.046	0.12*	0.13**	0.07	0.16**	0.16**	0.24**	0.24**	0.24**	0.31**	0.31**	0.31**	0.31**	0.31**	0.31**	0.31**	0.31**	1													
25 PERE_Performance	0.12**	0.042	0.15**	0.29**	0.022	0.055	0.09**	0.17*	0.12**	0.12**	0.12**	0.12**	0.12**	0.12**	0.12**	0.12**	0.12**	0.12**	0.12**	0.12**	0.12**	0.12**	0.12**	1												
26 PERE_Instructional Quality	0.092	0.031	0.095	0.078	0.027	0.09	0.027	0.09	0.027	0.09	0.027	0.09	0.027	0.09	0.027	0.09	0.027	0.09	0.027	0.09	0.027	0.09	0.027	0.09	1											
27 PERE_Resources Management	0.096	0.031	0.095	0.078	0.027	0.09	0.027	0.09	0.027	0.09	0.027	0.09	0.027	0.09	0.027	0.09	0.027	0.09	0.027	0.09	0.027	0.09	0.027	0.09	0.027	1										
28 PERE_Performance	0.096	0.031	0.095	0.078	0.027	0.09	0.027	0.09	0.027	0.09	0.027	0.09	0.027	0.09	0.027	0.09	0.027	0.09	0.027	0.09	0.027	0.09	0.027	0.09	0.027	0.09	1									
29 PERE_Performance (classroom)	0.096	0.031	0.095	0.078	0.027	0.09	0.027	0.09	0.027	0.09	0.027	0.09	0.027	0.09	0.027	0.09	0.027	0.09	0.027	0.09	0.027	0.09	0.027	0.09	0.027	0.09	0.027	1								
30 CLASS_Instructional Learning Formats	0.41**	0.29**	0.42**	0.48**	0.24**	0.43**	0.46**	0.50**	0.58**	0.49**	0.60**	0.25**	0.18**	0.35**	0.24**	0.49**	0.055	0.078	0.21**	0.20**	0.022	0.17**	0.17**	0.17**	0.17**	0.17**	0.17**	0.17**	0.17**	0.17**	0.17**	0.17**	0.17**	0.17**	0.17**	
31 ICALT_0-6th and remaining resources	0.29**	0.058	0.29**	0.28**	0.092	0.09	0.22**	0.19**	0.16**	0.26**	0.26**	0.62**	0.62**	0.62**	0.62**	0.62**	0.62**	0.62**	0.62**	0.62**	0.62**	0.62**	0.62**	0.62**	0.62**	0.62**	0.62**	0.62**	0.62**	0.62**	0.62**	0.62**	0.62**	0.62**	0.62**	
32 PERE_Instructional Quality	0.066	0.049	0.064	0.14**	-0.021	0.040	0.141**	0.17**	0.12**	0.16**	0.17**	0.47**	0.47**	0.47**	0.47**	0.47**	0.47**	0.47**	0.47**	0.47**	0.47**	0.47**	0.47**	0.47**	0.47**	0.47**	0.47**	0.47**	0.47**	0.47**	0.47**	0.47**	0.47**	0.47**	0.47**	
33 ICALT_0-6th and remaining pedagogy	0.19**	0.008	0.11**	0.17**	0.077	0.12*	0.20**	0.21**	0.14**	0.19**	0.20**	0.60**	0.60**	0.60**	0.60**	0.60**	0.60**	0.60**	0.60**	0.60**	0.60**	0.60**	0.60**	0.60**	0.60**	0.60**	0.60**	0.60**	0.60**	0.60**	0.60**	0.60**	0.60**	0.60**	0.60**	
34 CLASS_0-6th	0.76**	0.80**	0.75**	0.86**	0.52**	0.82**	0.80**	0.76**	0.76**	0.76**	0.76**	0.76**	0.76**	0.76**	0.76**	0.76**	0.76**	0.76**	0.76**	0.76**	0.76**	0.76**	0.76**	0.76**	0.76**	0.76**	0.76**	0.76**	0.76**	0.76**	0.76**	0.76**	0.76**	0.76**	0.76**	
35 ICALT_0-6th	0.20**	-0.018	0.20**	0.19**	0.11*	0.17**	0.19**	0.19**	0.19**	0.19**	0.19**	0.19**	0.19**	0.19**	0.19**	0.19**	0.19**	0.19**	0.19**	0.19**	0.19**	0.19**	0.19**	0.19**	0.19**	0.19**	0.19**	0.19**	0.19**	0.19**	0.19**	0.19**	0.19**	0.19**	0.19**	

* Correlation is significant at the .05 level
 ** Correlation is significant at the .01 level

model of the CLASS had the best but still unacceptable fit indices, $\chi^2(43) = 358.418$, $p < .001$, CFI = .897, RMSEA = .131.

6.1.2 ICALT and CETIT

Relatively speaking, the results indicated a poor model fit for the six-factor model of the *ICALT*, $\chi^2(449) = 2823.249$, $p < .001$, CFI = .558, RMSEA = .112, while all three CFA models of the *CETIT* suggested relatively better model fits than the *ICALT* ones. The eight-factor model of effective practices in the *CETIT*, $\chi^2(1091) = 4796.771$, $p < .001$, CFI = .618, RMSEA = .09, have better fit indices than those of the four-factor model of inspiring practices, $\chi^2(149) = 782.373$, $p < .001$, CFI = .659, RMSEA = .1, except for the SRMR. However, the full 12-factor model has overall the best fit (except for CFI) among all CFA models with $\chi^2(2144) = 802.596$, $p < .001$, CFI = .572, and RMSEA = .08.

7 Discussions

7.1 Teaching Effectiveness in Different Lens

This secondary analysis was intended to examine teacher effectiveness by comparing different classroom observation instruments. Theoretically, CLASS and ICALT have similar teaching dimensions, but our results showed that ICALT and CETIT were more closely correlated. We could not rule out that this closer relationship was a halo effect of the rater effect because the same raters rated them. While all three scales were reliable, some of the individual dimensions of ICALT and CETIT were internally inconsistent, contrary to the latest research (e.g., Ko & Li, 2020; Maulana et al., 2021). The most puzzling findings were the insignificant relationship of the factors in the confirmatory factor analyses of the three instruments in Table 16.6.

Table 16.6 Model fit indices of confirmatory factor models of CLASS, ICALT and CETIT

	CLASS			ICALT	CETIT		
	1-factor	2-factor	3-factor	6-factor	Effective-8 factor	Inspire-4 factor	12-factor
TLI	.716	.868	.679	.512	.593	.620	.546
CFI	.768	.897	.761	.558	.625	.659	.572
RMSEA	.186	.131	.205	.112	.095	.1	.08
SRMR	.105	.068	.154	.134	.113	.099	.112
χ^2/df	15.675	8.335	18.893	6.287	4.829	5.219	3.741

7.2 *Validity and Reliability of Instruments*

The major limitation of the current study was the poor validity and reliability of the instruments. Though the CLASS and ICALT have been validated in many international contexts, we failed to validate them in the selected sample. We do not intend to provide arguments for retaining the models with poor fit indices nor discuss strategies to modify the model to obtain an acceptable fit because this would go beyond the purpose of this paper. To our surprise, the two-factor model showed a better fit than the theoretical three-factor model. However, similar results were reported by Hafen et al. (2015), who found their bi-factor model fitted the MET data better than the original three-factor model. It is beyond this book chapter's scope to explore a possible revised three-factor model. Still, the results suggested that the CLASS could be inherently unstable because the *Instruction Support* domain is empirically more distinctive than the other domains.

Regarding instrument comparison, the CFA results favoured the CETIT slightly, more for its effective teaching component than its inspiring teaching component. Further studies on the relative significance of individual teaching dimensions (or subscales) will help us further teacher effectiveness research from scale or instrument development to teacher development conceptualising teaching practices ranked by difficulties (Ko et al., 2016).

We are also surprised that the reliability scores of some of the subscales of the ICALT and CETIT were unacceptably low. These results differed much from what we found in our previous projects (Ko et al., 2016, 2019a, b; Maulana et al., 2021). These results might raise concerns over the reliability of the raters' judgements. Given the high-inference nature of classroom observation instruments, ratings are expected to be evaluative. Though we had trained our raters and did calibration to minimise subjective biases in our observations, halo effects might affect the raters' judgements, making the results of the ICALT and CETIT more similar to each other than the CLASS. However, we are more inclined to suspect that this might be a side effect of a biased sample (see below). Still, further analyses to explore any rater effects seem wanting.

7.3 *Limitations with the Original MET Sample*

Conducting classroom observation or teacher evaluation research has been challenging because teacher evaluation is always a sensitive matter for practitioners. The MET lessons were not naturalistic and subject to self-selection bias because teachers and schools provided lesson video clips. There was little control over the quality of the recording and the settings. The video quality might affect the raters' judgments of student engagement as students were often off the screen. However, the secondary video data analysis could be a strength because this allowed other researchers to build up a video-based lesson database with other instruments.

Since we suspected there might be a problem using aggregate averages as references to select our lesson sample, we did another CFA with all the MET lessons to establish the scale validity, but the fit indices were also disappointing. We could not find any report concerning CLASS validation in the MET in its documentation or the literature. We could not identify what characteristics in the entire MET sample and our lesson sample might have caused the inadequate validation. Our assumption that the validations of the *ICALT* and *CETIT* were much affected by some unknown biased sample selection may not be justifiable as it seems. Moreover, we have not conducted further analyses to check systematic biases regarding teacher, school and district characteristics, as we assumed they would be marginal compared to variations in teaching quality.

7.3.1 Significance and Implications

Studying teachers' classroom practices and their effects is essential for teacher development and school improvement. We regard this study's significant implication in indicating the relative strengths and areas for teaching improvements (i.e., flexibility, innovative teaching, adjusting instructions and learner processing to inter-learner differences, teaching learning strategies). Future training on the *CETIT* and *ICALT* as reflective tools may benefit practitioners.

Despite the limitations discussed, this study provides data for instrument comparisons. Some teaching practices are comparable across instruments. Instrument comparison was already an essential focus in the MET project, which included six observation instruments, including more generic by nature, the *CLASS*, and more subject-specific ones like the *PLATO*, *MQI*, *QST*, and *UTOP*. Future research should extend comparisons to these subjects-specific instruments.

The secondary data analysis was a cost-effective strategy to connect two independent studies, the MET and *ICALT3* projects. The secondary data analysis could be done because the lessons were videotaped, providing opportunities to observe the same lessons again at different times and in research contexts. However, secondary data analysis is also limited by the quality of the original sample also limits secondary data analysis as the researchers who conduct secondary data analysis can do little to rectify flaws in the data collection processes.

8 Conclusion

It is tricky and controversial to define effective teaching or teaching effectiveness. Effective teaching requires criteria for effectiveness. The criteria implied in the various teaching dimensions in the *CLASS*, *ICALT* and *CETIT* refer to education objectives in general and teaching in particular. Visions about these criteria result from a political and societal debate, but educational professionals, teachers and schools can also participate in classroom observations. Going beyond identifying

effective classroom practice characteristics, we have uncovered the similarities and variations across teaching dimensions in different instruments. It was surprising that the CLASS could not be validated in our sample as in the original MET dataset. Despite limitations in the validity and reliability of the samples, we consider that our attempt to provide data for the ICALT3 project is at least partially fulfilled.

References

- 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.
- Bill & Melinda Gates Foundation. (2018). *Measures of effective teaching project: Frequently asked questions*. Retrieved from <http://k12education.gatesfoundation.org/blog/measures-of-effective-teaching-project-faqs/>
- Bacher-Hicks, A., Chin, M. J., Kane, T. J., & Staiger, D. O. (2019). An experimental evaluation of three teacher quality measures: Value-added, classroom observations, and student surveys. *Economics of Education Review*, 73, 101919.
- Boston, M., Bostic, J., Lesseig, K., & Sherman, M. (2015). A comparison of mathematics classroom observation protocols. *Mathematics Teacher Educator*, 3(2), 154–175.
- Cheng, Y. C. (1996). Total teacher effectiveness: new conception and improvement. *International Journal of Educational Management*, 10(6), 7–17.
- Cheng, Y. C., & Tsui, K. T. (1998). Research on total teacher effectiveness: Conception strategies. *International Journal of Educational Management*, 12(1), 39–47.
- Clark-Carter, D. (2005). Stanine Scores. In P. R. Rosenbaum, B. S. Everitt, & D. Howell (Eds.), *Encyclopaedia of statistics in behavioural science* (Vol. 4, p. 1893). Wiley.
- Clunies-Ross, P., Little, E., & Kienhuis, M. (2008). Self-reported and actual use of proactive and reactive classroom management strategies and their relationship with teacher stress and student behaviour. *Educational Psychology*, 28(6), 693–671.
- Danielson, C. (2013). *The framework for teaching evaluation instrument*. Danielson Group.
- Day, C., Sammons, P., Kington, E., & Regan, E. K. (2008). *Effective classroom practice: A mixed-method study of influences and outcomes*. ESRC.
- 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.
- Hafen, C. A., Hamre, B. K., Allen, J. P., Bell, C. A., Gitomer, D. H., & Pianta, R. C. (2015). Teaching through interactions in secondary school classrooms: Revisiting the factor structure and practical application of the classroom assessment scoring system–secondary. *The Journal of Early Adolescence*, 35(5–6), 651–680.
- Havik, T., & Westergård, E. (2020). Do teachers matter? Students' perceptions of classroom interactions and student engagement. *Scandinavian Journal of Educational Research*, 64(4), 488–507.
- Helmke, A., Schneider, W., & Weinert, F. E. (1986). Quality of instruction and classroom learning outcomes: The German contribution to the IEA classroom environment study. *Teaching and Teacher Education*, 2(1), 1–18.
- Hiebert, J., Gallimore, R., Garnier, H., Givvin, K. B., Hollingsworth, H., Jacobs, J., Chui, A. M., Wearne, D., Smith, M., Kersting, N., Manaster, A., Tseng, E., Etterbeek, W., Manaster, C., Gonzales, P., & Stigler, J. (2003). *Teaching mathematics in seven countries: Results from the TIMSS 1999 video study* (NCES (2003-013)). U.S. Department of Education, National Center for Education Statistics.

- 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., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalisability study. *Educational Researcher*, 41(2), 56–64.
- Jacobs, J. K., Kawanaka, T., & Stigler, J. W. (1999). Integrating qualitative and quantitative approaches to the analysis of video data on classroom teaching. *International Journal of Educational Research*, 31(8), 717–724.
- Kington, A., Sammons, P., Brown, E., Regan, E., Ko, J., & Buckler, S. (2014). *Effective classroom practice*. Open University Press.
- Ko, J. Y. O. (2010). Consistency and variation in classroom practice: A mixed-method investigation based on case studies of four EFL teachers of a disadvantaged secondary school in Hong Kong (Doctoral dissertation, University of Nottingham).
- Ko, J., & Li, W. (2020, April). *Effective teaching and inspiring teaching in different learning environments: Evidence from cluster analysis and SEM*. The 2020 Annual meeting of American Educational Research Association (AERA), Virtual Conference, USA.
- Ko, J., Chen, W., & Ho, M. (2015). *Teacher effectiveness and goal orientation* (School Report No. 1). Hong Kong Institute of Education.
- Ko, J., Sammons, P., & Kyriakides, L. (2016). *Developing instruments for studying inspiring teachers and their teaching practice in diverse contexts. Final report*. The Education University of Hong Kong.
- Ko, J., Fong, Y., & Xie, A. (2019a, August). *Effective and inspiring teaching in Math and Science classrooms: Evidence from systematic classroom observation and implications on STEM education*. WERA Annual Meeting in Tokyo.
- Ko, J., Sammons, P., Maulana, R., Li, W., & Kyriakides, L. (2019b, April). *Identifying inspiring versus effective teaching: how do they link and differ?*. The 2019 Annual meeting of American Educational Research Association (AERA), Toronto, Canada.
- Krammer, G., Pflanzl, B., Lenske, G., & Mayr, J. (2020). Assessing quality of teaching from different perspectives: Measurement invariance across teachers and classes. *Educational Assessment*, 26, 1–16.
- Learning Mathematics for Teaching Project. (2011). Measuring the mathematical quality of instruction. *Journal of Mathematics Teacher Education*, 14, 25–47.
- Lei, J., Chen, Z., & Ko, J. (2023). Differences in perceived instructional quality of the same classrooms with two different classroom observation instruments in China: Lessons learned from qualitative analysis of four lessons using TEACH and ICALT. In R. Maulana, M. Helms-Lorenz, & R. M. Klassen (Eds.), *Effective teaching around the world*. Springer.
- Liu, S., Bell, C. A., Jones, N. D., & McCaffrey, D. F. (2019). Classroom observation systems in context: A case for the validation of observation systems. *Educational Assessment, Evaluation and Accountability*, 31(1), 61–95.
- Maulana, R., Andre, S., Helms-Lorenz, M., Ko, J., Chun, S., & Shahzad, A. (2021). Observed teaching behaviour in secondary education across six countries: Measurement invariance and indication of cross-national variations. *School Effectiveness and School Improvement*, 32(1), 64–95.
- McCann, E., Riggall, A., Ingram, J., & Lindorff, A. (2020). *TALIS video study: Technical report*. DFE-RR109. Department for Education, UK.
- OECD. (2017). *TALIS 2018 video study and global video library on teaching practices*. OECD.
- 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. K., & Mintz, S. (2012). *Classroom Assessment Scoring System (CLASS) Manual, Secondary*. Teachstone.
- Piburn, M., & Sawada, D. (2000). Reformed teaching observation protocol (RTOP) reference manual. Technical Report. No. IN00-32. Arizona State University. <https://files.eric.ed.gov/fulltext/ED447205.pdf>

- Pöysä, S., Vasalampi, K., Muotka, J., Lerkkanen, M. K., Poikkeus, A. M., & Nurmi, J. E. (2019). Teacher-student interaction and lower secondary school students' situational engagement. *British Journal of Educational Psychology, 89*(2), 374–392.
- Roth, K. J., Druker, S. L., Garnier, H. E., Lemmens, M., Chen, C., Kawanaka, T., Rasmussen, D., Trubacova, S., Warvi, D., Okamoto, Y., Gonzales, P., Stigler, J., & Gallimore, R. (2006). *Teaching science in five countries: Results from the TIMSS 1999 video study* (NCES 2006-11). U.S. Department of Education, National Center for Education Statistics/U.S. Government Printing Office.
- Sammons, P., Kington, A., Lindorff-Vijayendran, A., & Ortega, L. (2014). *Inspiring teachers: Perspectives and practices*. CfBT Education Trust.
- Sammons, P., Lindorff, A. M., Ortega, L., & Kington, A. (2016). Inspiring teaching: Learning from exemplary practitioners. *Journal of Professional Capital and Community, 1*(2), 124–144.
- Schultz, S. E., & Pecheone, R. L. (2014). Assessing quality teaching in science. In *Designing teacher evaluation systems: New guidance from the measures of effective teaching project* (pp. 444–492). Jossey-Bass.
- Stigler, J. W., Gonzales, P., Kawanaka, T., Knoll, S., & Serrano, A. (1999). *The TIMSS videotape classroom study: Methods and findings from an exploratory research project on eighth-grade mathematics instruction in Germany, Japan, and the United States*. From <https://nces.ed.gov/pubs99/1999074.pdf>
- Stigler, J. W., Gallimore, R., & Hiebert, J. (2000). Using video surveys to compare classrooms and teaching across cultures: Examples and lessons from the TIMSS video studies. *Educational Psychologist, 35*(2), 87–10.
- Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). *Using multivariate statistics* (5th ed.). Pearson.
- Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education, 48*(6), 1273–1296.
- Taut, S., Jiménez, D., Puente-Duran, S., Palacios, D., Godoy, M. I., & Manzi, J. (2019). Evaluating the quality of teaching: Can there be valid differentiation in the middle of the performance distribution? *School Effectiveness and School Improvement, 30*(3), 328–348.
- van de Grift, W. (2007). Quality of teaching in four European countries. *Educational Research, 49*, 127–152.
- van de Grift, W. (2014). Measuring teaching quality in several European countries. *School Effectiveness and School Improvement, 25*(3), 295–311.
- Virtanen, T. E., Lerkkanen, M. K., Poikkeus, A. M., & Kuorelahti, M. (2015). The relationship between classroom quality and students' engagement in secondary school. *Educational Psychology, 35*(8), 963–983.
- Walkington, C., & Marder, M. (2014). Classroom observation and value-added models give complementary information about quality of mathematics teaching. In T. Kane, K. Kerr, & R. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the measuring effective teaching project* (pp. 234–277). Wiley.
- Walkington, C., & Marder, M. (2018). Using the UTeach observation protocol (UTOP) to understand the quality of mathematics instruction. *ZDM Mathematics Education, 50*, 507–519.
- Wallace, T. L., Parr, A. K., & Correnti, R. J. (2020). Assessing teachers' classroom management competency: A case study of the classroom assessment scoring system—secondary. *Journal of Psychoeducational Assessment, 38*(4), 475–492.
- World Bank. (2019). *TEACH manual*. World Bank. https://www.worldbank.org/en/topic/education/brief/teachhelping-countries-track-and-improve-teaching-quality?cid=EXT_WBEmailShare_EXT
- Zhao, Y., & Ko, J. (2022, Apr). *Classroom orchestration as a potent stimulator of dialogic teaching and learner engagement in higher vocational classrooms*. [Paper presentation] American Educational Research Association (AERA 2022) Annual Meeting (Virtual), San Diego, California.

Dr. James Ko is an Associate Professor at the Department of Policy Leadership and Co-Director of the Joseph Lau Luen Hung Charitable Trust Asia Pacific Centre for Leadership and Change at the Education University of Hong Kong. Before his doctoral study, James was an EFL teacher for about 20 years and led two functional teams in a secondary school for 10 years. He is a recurrent grantee of the RGC and UGC grants and the principal investigator of 23 projects, collaborating with local academics and overseas researchers on 40 projects. He has supervised 14 doctoral students with 8 completed.

Zhijun Chen is a Doctoral Researcher at the Department of Education at the University of Bath (UK). She holds an MSc in Psychology from the University of St Andrews (UK). She also works as a research assistant with Dr James Ko at the Education University of Hong Kong, focusing on teaching quality and teaching assessment in multiple cultures. Her research interests include educational effectiveness, large-scale international assessments (PISA, TIMSS, PIRLS, ERCE, etc.), education inequality, and classroom observation.

Celia Lei is a lecturer at Shaoyang University in Mainland China and a doctoral candidate at the Education University of Hong Kong, where she graduated with an MEd degree and worked as a research assistant at the Department of Educational Policy and Leadership. She published on teaching quality of public schools in an inland province in Mainland China and has extensive research experience using various classroom observation instruments. Her research interests include dialogic teaching, teaching quality, and metacognitive teaching.

Ridwan Maulana is an associate professor at the Department of Teacher Education, University of Groningen, the Netherlands. His major research interests include teaching and teacher education, factors influencing effective teaching, methods associated with the measurement of teaching, longitudinal research, cross-country comparisons, effects of teaching behaviour on students' motivation and engagement, and teacher professional development. He has been involved in various teacher professional development projects including the Dutch induction programme and school-university-based partnership. He is currently a project leader of an international project on teaching quality involving countries from Europe, Asia, Africa, Australia, and America. He is a European Editor of Learning Environments Research journal, a SIG leader of Learning Environments of American Educational Research Association, and chair of the Ethics Commission of the Teacher Education.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

