Chapter 5

A Probabilistic Model for Feedback on Teachers' Instructional Effectiveness: Its Potential and the Challenge of Combining Multiple Perspectives

Rikkert van der Lans

Abstract This chapter describes research into the validity of a teacher evaluation framework that was applied between 2012 and 2016 to provide feedback to Dutch secondary school teachers concerning their instructional effectiveness. In this research project, the acquisition of instructional effectiveness was conceptualized as unfolding along a continuum ranging from ineffective novice to effective expert instructor. Using advanced statistical models, teachers' current position on the continuum was estimated. This information was used to tailor feedback for professional development. Two instruments were applied to find teachers' current position on the continuum, namely the *International Comparative Assessment of Learning and Teaching* (ICALT) observation instrument and the *My Teacher–student questionnaire* (MTQ). This chapter highlights background theory and central concepts behind the project and it introduces the logic behind the statistical methods that were used to operationalize the continuum of instructional effectiveness. Specific attention is given to differences between students and observers in how they experience teachers' instructional effectiveness and the resulting disagreement in how they position teachers on the continuum. It is explained how this disagreement made feedback reports less actionable. The chapter then discusses evidence of two empirical studies that examined the disagreement from two methodological perspectives. Finally, it makes some tentative conclusions concerning the practical implications of the evidence.

Keywords Teacher evaluation · Teaching quality · Measures · Teacher development · Feedback

1 Introduction

Tess is a school leader who must decide how to spend the school's resources on professional development. Today she has an assessment interview with John of whom

R. van der Lans (✉)
LUMC-Curium - Child and Adolescent Psychiatry, Leiden, The Netherlands
e-mail: r.m.van_der_lans@lumc.nl

she has just visited a lesson yesterday. Her rubric scorings signal some clear directions for improvement in the clarity of John's front class explanations, yet the student questionnaire administered one month ago gives few signs of poor explanation skills. Instead, the student results signal that John could improve on the interactivity of his instructions. Tess wants to use the interview to plan and guide John's further professionalization. However, how can she use the available information to provide John with actionable feedback that likely adds to John's teaching? Also, Tess knows that John wants to discuss opportunities to participate in further training. Yet, the evidence of John's instructional skill is inconclusive. Hence, on what grounds should she accept or refuse the request?

Imaginary situation based on conversations with school leaders and teachers. The last decade teacher evaluation has had a central position in policies aiming to improve educational quality in many countries (e.g., Doherty & Jacobs, 2013; Isoré, 2009; Nusche et al., 2014). In the Dutch context, the Ministry of Education published the "teacher agenda" which documented several challenges, objectives, and policy measures meant to increase the quality of the Dutch teacher workforce. One objective was to increase the frequency of performance evaluations in schools (Nusche et al., 2014; OECD, 2016). In the eyes of the policy-makers, performance evaluations were a means to turn schools into "learning organizations" by functioning as a yearly update and a reminder of teachers' and school leaders' commitment to increase educational quality. To realize this objective, the councils for primary (PO-raad) and secondary (VO-raad) education and the teacher labor unions together agreed to install a new differentiated payment system and to assign every teacher a personal professionalization budget (Nusche et al., 2014; OECD, 2016). This created an incentive for teachers to request for a performance evaluation interview to discuss evidence of instructional effectiveness. If evidence of effectiveness was insufficient to qualify for a salary raise, the teacher should be informed about steps and/or skills required to qualify. Teachers could use their personal professionalization budget to train these skills. In practice these policies implied that school leaders, like Tess, were confronted with the task to distinguish between "average", "good" and "excellent" teachers, to give teachers feedback about what they needed to learn, and to organize the conditions under which teachers can start to learn.

The research project described in this chapter took place within this context and examined statistical methods and models that could assist school leaders to distinguish between teachers in terms of their level of instructional effectiveness. Furthermore, the new methods needed to result in feedback that clearly indicated a specific direction for improvement. Instruments used to collect data were student questionnaires and classroom observations. These two instruments were chosen because they share the strength that they collect direct observations of teachers' classroom behavior (Darling-Hammond, 2013; Goe et al., 2008; Peterson, 2000). However, as the situation of Tess and John shows, the feedback resulting from the student questionnaire and the feedback resulting from the classroom observation instrument do not always agree.

The chapter starts with an introduction of central concepts and how these were operationalized. The applied methods are introduced at a conceptual level and it is discussed how the models relate to other statistical methods that are commonly applied. After the background section, the chapter focuses on the problem of agreement between feedback sampled with student and classroom observation instruments.

2 Background Theory and Definitions of Central Concepts

In the sketch at the beginning of this chapter, Tess is wondering how she may use student questionnaires and classroom observation instruments as two complementary instruments to provide John with actionable feedback. This highlights some central concepts of this chapter, namely instructional effectiveness, improvement, and actionable feedback.

2.1 *Instructional Effectiveness*

In this chapter, instructional effectiveness is viewed as an estimation of the degree to which teachers' classroom behavior is expected to give students the opportunity to maximize their learning potential. By stating that instructional effectiveness provides students with opportunities to learn it is clarified that instructional effectiveness is associated with, but not identical to, student achievement and school success which are realizations of these opportunities. Furthermore, the definition clarifies that instructional effectiveness is estimated meaning that any claim about it is surrounded by some level of uncertainty. The research described in this chapter has operationalized instructional effectiveness using two instruments, namely the International Comparative Assessment of Learning and Teaching (ICALT)—which is a classroom observation instrument—and the My Teacher questionnaire—which is a student questionnaire. The ICALT and My Teacher questionnaire instruments conceptualize an effective instructor as a teacher that scores high on six domains of instruction. The six domains are labeled “safe and stimulating learning climate”, “efficiency of classroom management”, “clear and structured explanations”, “intensive and interactive instructions”, “teaching students learning strategies”, and the “adaptation of instructions to individual student needs”. Table 1 details the conceptualizations of each domain. Several studies from different countries provide evidence suggesting that the items included in the ICALT and My Teacher questionnaire cluster according to these six domains (André et al., 2020; Maulana & Helms-Lorenz, 2016; van de Grift et al., 2011).

Table 1 An overview of the six domain and their conceptualization

Domain	Conceptualization
• Safe and stimulating learning climate	A safe and stimulating learning climate is established when the teacher and students trust and respect each other.
• Efficient classroom management	An efficient classroom management is structured by clear procedures, routines, and rules about where and how learning takes place.
• Clear and structured explanations	Clear and structured explanations prompt students' prior knowledge, emphasize critical knowledge, and regularly checks on students' comprehension of content.
• Intensive and interactive instructions	Intensive and interactive instructions stimulate teacher–student and student–student interaction by questioning, collaborative group work, having students explain topics to one another, or asking students to think aloud.
• Teaching students learning strategies	Teaching students learning strategies enhance students' metacognitive skills and self-regulated learning.
• Adaptation of instructions to individual student needs	Adaptation of instruction means that teachers adjust their instructional practice to specific students' learning needs by, for example, allowing flexible time to complete assignments or providing additional explanation to small groups.

2.2 Improvement

Another central term in this chapter is improvement, which suggests that teachers can learn or be trained to become more effective instructors. Analogous to Berliner's (2004) novice-expert continuum, the research project and evidence discussed here conceptualizes the improvement of instructional effectiveness as unfolding along a continuum ranging from completely ineffective instruction to completely effective instruction. To illustrate how we operationalized this continuum we start with a one-item example, "This teacher uses time efficiently". Research suggests that more effective instructors in general use time more efficiently than less effective instructors (Muijs et al., 2014). The x -axis in Fig. 1 visualizes the continuum of effective instruction. The y -axis represents the probability on a positive item score. In line with the above statement, Fig. 1 predicts that highly effective instructors have near 100% probability on positive scores and that low effective instructors have near 0% probability. Furthermore, only when teachers have acquired a certain level of instructional effectiveness, are they predicted to start to learn how to use time efficiently. This can be inferred from Fig. 1 by observing that the probability on positive scores starts to rise at a certain location on the continuum. Key to the perspective taken in this

Fig. 1 Increase in probability on a positive response to the classroom observation item “This teacher uses time efficiently”

chapter is that training of teachers’ instructional effectiveness is optimal when it is focused on items that match the teacher’s location on the continuum.

2.2.1 Relation of the Applied Relatively Novel Statistical Model to Other Statistical Models

The proposition that all items measure instructional effectiveness are associated with a single continuum seems conflicting with research that groups items of instructional effectiveness according to dimensions of teaching quality which load on separate (statistical) factors (e.g., studies applying factor analysis). Figure 2 is used to discuss and visualize the relationship between the continuum discussed in this chapter and factor analysis results. Figure 2 again visualizes the continuum of instructional effectiveness, but now includes multiple items. Solid, dashed, and dotted lines indicate clusters of items that have high(er) inter-item correlations (i.e., load on separate factors). The reader can move the icon directly below the Figure to logically derive this. For example, teachers positioned at the icon have an approximately 50% probability on positive scores on the dotted items, but near 0% probability on positive scores on the dashed and solid ones. When the teacher moves up the continuum, the probability on positive scores on the dashed items increases first, while the probability on the dotted items remains high and the solid items remains low. Thus, some teachers likely have high scores on dotted items and low on all other items, other teachers likely have high scores on the dotted and high on the dashed items but low on the solid items, and yet others likely have high scores on all items. However, it is unlikely that teachers score high on the solid but low on the other items. This scoring pattern, which in factor analytic literature is referred to as the simplex pattern, is

Fig. 2 The continuum of instructional effectiveness in which indicators are grouped into three factors

detected by factor analysis as a sign that the dotted items and dashed items are in distinct clusters (interested readers may consult Browne [1992] or Jöreskog [1978] for further details). It is acknowledged that the Figure presents an oversimplification of the relationship of factor analysis with the continuum described in the chapter. For example, item slopes, i.e., the steepness with which the s-curved lines increase, are rarely exactly parallel. Such differences in item slope also impact on the assignment of items to factors. However, Fig. 2 illustrates the basic rationale of how different factors on a single continuum.

2.2.2 The Sequence of Clusters Along the Continuum

The research group at the University of Groningen has given much empirical attention to the ordering of these factors along the continuum of instructional effectiveness (e.g., Maulana et al., 2015a, 2015b; van de Grift et al., 2011, 2014; van der Lans et al., 2015, 2018, 2021). The results indicated an ordering of the factors in the sequence in which the factors are presented in Table 1, thus: (1) safe and stimulating learning climate, (2) efficient classroom management, (3) clear and structured explanations, (4) intensive and interactive instructions, (5) teaching students learning strategies, and (6) adaptations of instructions to the individual students' learning needs. The validity of this ordering was further corroborated by other research in Cyprus that applied the same statistical models to their own questionnaire and observation instruments and which reported broadly similar results (e.g., Kyriakides et al., 2009, 2018).

Based on these results, the author developed a feedback report to provide teachers with information of our best estimate of their current position on the continuum of instructional effectiveness and our best estimate of what the teacher could improve

on next. Figure 3 presents two reports that were applied by the author to give teacher feedback. The left report concerns the classroom observation instrument and the right concerns the student questionnaire instrument. The reports show a table with three columns. In the column “level” are the six identified levels (or domains) of instructional effectiveness. The column “item” lists the items included in the instruments. Finally, the column, “teacher score” indicates what probably went well (darkest grey top area), what probably can be learnt next (lightest grey middle area), and what probably is beyond the teachers’ competency to learn yet (grey lowest area). The Asterix indicates the exact teacher position on the continuum of instructional effectiveness.

2.3 Actionable Feedback

The third central concept in this chapter is actionable feedback. Cannon and Witherspoon (2005) describe actionable feedback as feedback that leads to learning and increased performance. Evidence indicates that feedback is actionable when, (1) it is directed at the task—not the person, (2) it is unambiguous and specific, and (3) it has clear implications for action (Cannon & Witherspoon, 2005; Kluger & DeNisi, 1996). The aim was to design feedback reports such that they would assist the feedback giver, which presumably is a school leader or coach, to be actionable. Therefore, the reports emphasize on what the teacher does (e.g., the teacher involves students, explains clearly), it attempts to communicate as specific as possible about what went well and what can be improved. In addition, reports accompanied this information with specific implications that would now require action. Participating teachers found this approach informative and often recognized themselves. Nonetheless, the actionability of the feedback was hindered by various organizational and psychometric factors. Organizationally, schools lacked an infrastructure to support training activities at this level of precision. Though, this was not part of the research project discussed here, it is nonetheless important to mention. Psychometrically, the disagreement between students and observers created uncertainty about the reliability of the estimates. Take, for example, the two feedback reports in Fig. 3. The students on average positioned the teacher on the item “my teacher involves me in the lesson” and signal that the teacher should focus to improve on the clarity and structuredness of explanation. The classroom observer, however, positioned the teacher on the item “encourages students to apply what they have learnt” and signals that the teacher should improve on teaching students learning strategies. Moreover, the observer’s report suggests no problems with the clarity and structuredness of the teacher’s explanations. Hence, in case of disagreement the feedback reports no longer unambiguously communicate what went well and what domains of instruction needed improvement. Also, the implications for action were no longer clear.

Teacher:	#Name	Teacher:	#Name
Class:	#Name	Class:	#Name
School:	#Name	School:	#Name
Level	Item:	Level	Item:
	This teacher...		My teacher...
safe learning climate	shows respect for students in behavior and language	Safe Learning Climate	approaches me with respect.
safe learning climate	creates a relaxed atmosphere	Safe Learning Climate	helps me if I do not understand
efficient classroom management	ensures effective class management	Efficient Classroom Management	answers my questions.
safe learning climate	supports student self-confidence	Efficient Classroom Management	prepares his/her lessons well.
clear and structured explanation	explains the subject matter clearly	Efficient Classroom Management	makes sure that I know what to do.
efficient classroom management	ensures that the lesson runs smoothly	Clear and Structured Explanation	explains how I need to do things.
clear and structured explanation	gives well-structured lessons	Safe Learning Climate	makes clear what I need to learn for a test.
efficient classroom management	uses learning time efficiently	Safe Learning Climate	makes sure that I treat others with respect.
clear and structured explanation	clearly explains teaching tools and tasks	Efficient Classroom Management	states clearly when assignments/tasks are due.
safe learning climate	ensures mutual respect	Clear and Structured Explanation	makes sure that I pay attention.
efficient classroom management	checks during processing whether students are carrying out tasks properly	Clear and Structured Explanation	makes sure that I use my answers are good or not.
clear and structured explanation	gives feedback to students	Efficient Classroom Management	uses clear examples.
clear and structured explanation	involves all students in the lesson	Safe Learning Climate	pays attention to me.
clear and structured explanation	encourages students to do their best	Clear and Structured Explanation	explains everything clearly to me.
clear and structured explanation	checks during instruction whether students have understood the subject matter	Clear and Structured Explanation	involves me in the lesson.
intensive and interactive instructions	uses teaching methods that activate students	Clear and Structured Explanation	makes sure that I do my best.
intensive and interactive instructions	provides interactive instruction	Efficient Classroom Management	explains the purpose of the lesson clearly.
intensive and interactive instructions	asks questions that encourage students to think	Intensive and Interactive Instruction	makes sure that I use my time effectively.
intensive and interactive instructions	encourages students to reflect on solutions	Clear and Structured Explanation	encourages me to think.
intensive and interactive instructions	explains the lesson objectives at the start of the lesson	Efficient Classroom Management	states the lesson objectives.
intensive and interactive instructions	has students think out loud	Clear and Structured Explanation	repeats what we have learnt in the previous lesson.
intensive and interactive instructions	boosts the self-confidence of weak students	Efficient Classroom Management	makes sure that I keep on working.
teaching learning strategies	encourages students to think critically	Intensive and Interactive Instruction	motivates me to think.
teaching learning strategies	teaches students how to simplify complex problems	Adaptation of Instructions	makes connections to what I already know.
teaching learning strategies	encourages students to apply what they have learned	Intensive and Interactive Instruction	asks me questions that I need to think about.
adaptation of instructions	checks whether the lesson objectives have been achieved	Intensive and Interactive Instruction	takes into account what I already know.
adaptation of instructions	adapts processing of subject matter to student differences	Adaptation of Instructions	motivates me.
teaching learning strategies	encourages the use of checking activities	Intensive and Interactive Instruction	stimulates me to cooperate with my classmates.
teaching learning strategies	asks students to reflect on approach strategies	Adaptation of Instructions	checks whether I have understood the content of the lesson.
teaching learning strategies	teaches students to check solutions	Intensive and Interactive Instruction	talks interestingly.
adaptation of instructions	offers weak students additional learning and instruction time	Teaching Learning Strategies	makes me feel self-confident with difficult tasks.
adaptation of instructions	adapts instruction to relevant student differences	Teaching Learning Strategies	teaches me to check my solutions.
		Teaching Learning Strategies	tells how I should learn something.
		Teaching Learning Strategies	knows what I have difficulty with.
		Teaching Learning Strategies	lets me explain to him/her how I tackled the task/assignment.
		Teaching Learning Strategies	lets me explain the content of the lesson to other students.
		Teaching Learning Strategies	lets me summarize the content of the lesson.
		Teaching Learning Strategies	asks me how I am going to learn the content of the lesson.

Fig. 3 ICALT (left) and My Teacher (right) feedback reports. Darkest-grey = probably skilled (top area), light grey (middle area) = try to learn now, grey (lowest area) = probably unskilled in (Note These two feedbacks are slightly different versions of the ones that were originally reported to teachers)

3 Prior Research on the Disagreement Between Classroom Observation and Student Questionnaires

Prior research suggests that the disagreement between students and observers may be frequent and/or substantial. Studies documenting correlations between observation and survey measures mostly report modest correlations in the range of 0.15–0.30 (e.g., De Jong & Westerhof, 2001; Ferguson & Danielson, 2014; Howard et al., 1985; Martínez et al., 2016; Maulana & Helms-Lorenz, 2016). Designs varied considerably between studies, however. For example, De Jong and Westerhof (2001) report on the correlation between a classroom observation instrument and a student questionnaire that had considerably different factor structure, whereas Maulana and Helms-Lorenz report on the correlation of the ICALT and My Teacher questionnaire that have an overlapping factor structure. Because the study by Maulana and Helms-Lorenz (2016) applied the same instruments, their results are most relevant to the discussion in this chapter. They report a correlation of 0.26. This correlation was replicated in the data that was used in the research project that is reported on in this chapter. This modest correlation suggests that feedback reports of students and observers will more often disagree than agree. An exception to the above list of studies reporting modest correlations is the study by Murray (1983), who reports a correlation of 0.76. We will return on Murray's study somewhat later in this chapter.

4 Studying Evidence of Agreement and Disagreement Between Questionnaires and Classroom Observation Instruments

Two perspectives can be taken to compare the feedback reports presented in Fig. 3. The first perspective focuses on the teacher's position and as we have seen this leads to the conclusion that the students and observers disagree. The alternative perspective focuses on the ordering of the items and domains on the continuum of instructional effectiveness. From this perspective the students and observers mostly agree. Both the classroom observation and student questionnaire feedback report start with items related to safe and stimulating learning climate and end with items related to teaching students learning strategies and adaption of instruction to individual students' learning needs. The only two domains that are ordered differently by the two methods are the just mentioned final two domains.

Van der Lans et al. (2019) went one step further and showed that the My Teacher–student questionnaire and the ICALT classroom observation instrument items can be concurrently calibrated on the same continuum of instructional effectiveness. Table 2 lists the joint item ordering mixing observation and questionnaire items. Items denoted with an “s” are student questionnaire items and items denoted with an “o” are classroom observation items.

Table 2 Item ordering that resulted from the concurrent calibration of ICALT observation and My Teacher questionnaire items. This table was originally published in van der Lans et al. (2019) (O = ICALT observation item; S = My teacher questionnaire item)

Level	Item	Description: This/My teacher...
Climate	O1	Shows respect for students in behavior and language
Climate	S21	Treats me with respect
Climate	O2	Creates a relaxed atmosphere
Management	S20	Prepares his/her lesson well
Management	O7	Ensures effective class management
Climate	O3	Supports student self-confidence
Climate	S40	Helps me if I do not understand
Clear explanation	O9	Explains the subject matter clearly
Climate	S6	Answers my questions
Management	O5	Ensures that the lesson runs smoothly
Climate	O4	Ensures mutual respect
Clear explanation	O14	Gives well-structured lessons
Management	S3	Makes clear what I need to study for a test
Management	O8	Uses learning time efficiently
Management	S19	Makes clear when I should have finished an assignment
Climate	S8	Ensures that I treat others with respect
Climate	S1	Ensures that others treat me with respect
Clear explanation	S13	Explains the purpose of the lesson
Clear explanation	S24	Uses clear examples
Management	S23	Ensures that I pay attention
Management	S26	Applies clear rules
Management	O6	Checks during processing whether students are carrying out tasks properly
Management	S2	Ensures that I use my time effectively
Clear explanation	O15	Clearly explains teaching tools and tasks
Clear explanation	O10	Gives feedback to students
Clear explanation	O11	Involves all students in the lesson
Clear explanation	S39	Involves me in the lesson
Clear explanation	O13	Encourages students to do their best
Clear explanation	S33	Ensures that I know the lesson goals
Interactive instruction	S17	Encourages me to think for myself
Interactive instruction	O19	Asks questions that encourage students to think
Interactive instruction	S12	Ensures that I keep working
Interactive instruction	O16	Uses teaching methods that activate students
Interactive instruction	S30	Stimulates my thinking

(continued)

Table 2 (continued)

Level	Item	Description: This/My teacher...
Interactive instruction	O21	Provides interactive instruction
Clear explanation	O12	Checks during instruction whether students have understood the subject matter
Interactive instruction	O20	Has students think out loud
Differentiation	S25	Connects to what I am capable of
Differentiation	S34	Checks whether I understood the subject matter
Learning strategies	O30	Encourages students to apply what they have learned
Learning strategies	S16	Teaches me to check my own solutions
Learning strategies	O31	Encourages students to think critically
Differentiation	S36	Knows what I find difficult
Differentiation	O23	Checks whether the lesson objectives have been achieved
Learning strategies	O28	Encourages the use of checking activities
Learning strategies	O29	Teaches students to check solutions
Differentiation	O25	Adapts processing of subject matter to student differences
Differentiation	O26	Adapts instruction to relevant student differences

Studying Table 2 teaches us that similarly phrased questionnaire and observation items occasionally have similar positions on the continuum. Examples are S39 “my teacher involves me in the lesson” (student) and O11 “this teacher involves all students in the lesson” (classroom observation) and: S17 “my teacher encourages me to think for myself” (student) and O19 “this teacher asks questions that encourage students to think” (classroom observation). However, the questionnaire and classroom observation instrument also contained several items that were instrument unique, but which nevertheless could be calibrated on the same continuum. The considerable overlap between the questionnaire and observation instrument has clear practical implications. Suppose that two reports in Fig. 3 would locate a teacher on items related to the same domain—e.g., both on the domain clear and structured explanation—then the observation and student feedback reports give identical suggestions for improvement and, thus, would be more actionable. That is, there is no thinkable scenario in which feedback reports are actionable and in which the ordering of teaching behaviors (items) on the continuum varies between the methods.

The combination of agreement in item ordering and disagreement in teacher location also has theoretical implications, because these findings do not fit well with most prior beliefs, theory, and hypotheses concerning the disagreement between questionnaire and observation instruments. For example, a long-standing tradition in educational psychology studies biases in instrument scores. Bias is generally examined by regressing the teacher scores on variables other than instructional effectiveness. Studies in the MET project, like Martínez et al. (2016), regressed the “teacher scores” on variables that were hypothesized to bias measurement. This resulted in a set of ‘bias-corrected’ teacher scores related to the student questionnaire and a set

of ‘bias-corrected’ teacher scores related to the classroom observation instrument. These corrected scores were then correlated. However, the resulting correlations were similar to the correlations reported in studies that do not correct for bias (cf. Maulana & Helms-Lorenz, 2016; Martínez et al., 2016). More in general, hypotheses reflecting the belief that inferences based on scores need to be corrected for bias do not fit well with the evidence discussed so far. When scores obtained with the instruments are biased and not indicative of instructional effectiveness, then how can we explain the high similarity in the item ordering along the continuum.

The evidence also is difficult to align with another prominent hypothesis, namely the perspective-specific validity hypothesis stated by Kunter and Baumert (2006). Kunter and Baumert proposed that, despite disagreement, scores obtained with distinct instruments can be used to make valid inferences about teachers’ instructional effectiveness, given that the instruments are well-designed and administered. Kunter and Baumert do not clearly define what they mean with “perspectives” and this makes it complex to empirically assess their hypothesis (see also Fauth et al., 2020). However, many seem to understand different perspectives as meaning that some instruments might be more sensitive to tap certain aspects of instructional effectiveness. The difference in sensitivity explains the modest correlation. Also, using multiple instruments could help to offset blind spots thereby allowing for a fuller and richer picture of instructional effectiveness. The current evidence is insufficient to completely verify this idea, but an analysis of the unique items in Table 2 provides surprisingly limited support for it. Take, for example, the item S3 “my teacher makes clear what I need to study for a test”. Despite that this item mentions unique content (“what students need to study for a test” is not part of the ICALT observation list because observers usually cannot know this), the item S3 does not have a unique position on the continuum. We could leave out item S3 without losing much information about teachers’ instructional effectiveness. As another example, take item S34 “my teacher checks whether I understood the subject matter”. The phrasing “whether I understood” focuses on the individual student and such focus is not included in any of the classroom observation instrument items. Nonetheless, item S34 is very closely located to the item O12 “this teacher checks during instruction whether students have understood the subject matter” which measures the same content, but has observers focus on the “average” student in the class. In sum, the evidence provides no strong indications that differences between instruments in terms of item content and item focus can explain disagreement in how students and observers position teachers on the continuum of teaching effectiveness.

The disagreement in teachers’ position on the continuum was examined in another study. Central in that study was the hypothesis that disagreement in teachers’ position on the continuum changes as a function of measurement reliability (van der Lans, 2018). Central in the study were two claims of which the correctness was empirically assessed. First, it was claimed that the scores assigned by observers reflect the average student in the class. Therefore, agreement was expected to increase when classroom observation scores were correlated with class average questionnaire score, instead of scores assigned by a single student. Secondly, it was claimed that student responses to questionnaire items reflect the teachers’ typical teaching across

many lessons. Therefore, the agreement was expected to increase when classroom observation scores sampled from different lessons were averaged. The study applied generalizability theory to test these predictions and found support for both of them. The predicted correlation is lowest when scores of a single student's questionnaire are correlated with one classroom observation score of one single lesson and the more student questionnaires are sampled the higher the predicted correlation with the observation score of a single lesson becomes. Also, the predicted correlation increases when the observation scores are aggregated over multiple lessons, and, again, the more lessons are sampled the higher the predicted correlation. The study results suggest that the correlation between questionnaire and classroom observation instruments increases to 0.76 when the classroom observation scores concern an aggregate of seven different lesson visits performed by seven different observers and when the student questionnaire is administered in the same class and spans scores of 25 different students. This correlation of 0.76 was interesting because of its correspondence to the correlation reported by Murray (1983), which was also 0.76. Murray estimated that correlation based on the aggregate classroom observation score of six to eight lesson visits by three different observers and a student questionnaire administered in the same class and which score was aggregated over all students in the class. The increase in the expected correlation between the questionnaire and classroom observation instrument is graphically presented in Fig. 4. The y-axis in Fig. 4 gives the predicted correlation. The x-axis indicates the number of students in the class. The separate lines indicate how predictions differ when the number of classroom

Fig. 4 Predicted increase in correlations (ρ) between the MTQ student questionnaire and ICALT classroom observation instrument for an increasing number of administered classroom observations. The correlations apply when questionnaires and classroom observations are performed within the same class and span no more than one school year. Van der Lans (2018) reports predictions related to other situations (e.g., questionnaires and observations spanning different classes)

observers and lesson moments sampled with the classroom observation instrument varies.

The implications of the results in Fig. 4 are not yet well understood. There are varying possible interpretations. One interpretation is that more valid inferences are made with the questionnaire and classroom observation instruments when scores are aggregated over many students and over many lessons and observers, respectively. This interpretation aligns well with studies suggesting that the reliability of single student questionnaire scores is unreliable (Marsh, 2007) and that classroom observations of one single lesson are unreliable snap-shots (Hill et al., 2012; Praetorius et al., 2014; van der Lans et al., 2016). This interpretation has considerable implications for the number of classroom observations and student questionnaires that need to be administered at schools. However, this interpretation does not align well with the rationale behind the hypotheses of van der Lans (2018). That is, the only reason why it is predicted that the correlation between the classroom observation instrument and the student questionnaire increases as a function of the number of lesson visits included in the aggregate is that the students are also expected to aggregate their experiences across many lessons when scoring the questionnaire. If questionnaires would be able to tap students' experiences about one particular lesson, then it would be predicted that the correlation is highest when the questionnaire results are correlated with classroom observations concerning that same particular lesson. This study was unable to empirically examine this, however. Similarly, the finding that questionnaire scores obtained with single students have low correlation with the classroom observation scores is, in the study by van der Lans (2018), explained by the assumption that classroom observers score the instructional effectiveness towards the "average" student. If observers would have been instructed to score the classroom observation items in relation to one particular student, the correlation is predicted to be highest for the particular student-observer dyad (compared to other dyads). Again, the study was unable to empirically examine this claim.

5 Discussion and Conclusion

5.1 *Potential Implication Teacher Evaluation in Schools*

Based on the above discussions, what can we advise Tess and John? What we can say is that the evidence so far suggests that single classroom visits rarely will show agreement with a single administration of the student questionnaire. Also, the evidence generally indicates that this has little to do with the interpretation of item content by students and observers. The item ordering estimated across all students is very similar to the item ordering across all observers. This does not imply that the quality of the item phrasing is unimportant, however. The evidence indicates that when items are well-formulated the students can score items related to the same domains of instructional effectiveness very similar. We might advise Tess to postpone the

performance evaluation interview and administer some more classroom observations. The evidence presented in this chapter indicates that this increases the chance on agreement. However, it is not always possible to schedule additional classroom observations. Alternatively, we might advise Tess and John to focus on one result. Perhaps John wants to improve on his instructional effectiveness when teaching certain subject matter and because the classroom observation visit took place when John was teaching this particular subject matter, the classroom observation results are favored over the student questionnaire results. However, while this last advice might be intuitive to some, we must acknowledge that it is full of untested claims and hypotheses.

5.2 *What to Do Next?*

One direction for future research concerns the construction of student questionnaires that can help us to make valid inferences about the instructional effectiveness of single lessons. The Impact! tool might be a potential example of such an instrument (Bijlsma et al., 2019). The alternative item phrasing of the Impact! questionnaire provides opportunities to assess the hypothesis that correlations between student questionnaires and classroom observation instruments are attenuated because students aggregate their experiences over many lessons when scoring regular questionnaire items.

Another direction for future research concerns the commonly shared understanding among researchers that some instruments have higher sensitivity to measure certain aspects/behaviors of instructional effectiveness and that using multiple instruments could help to offset blind spots. When instruments have blind spots concerning the measurement of instructional effectiveness, then we would expect “gaps” in the continuum of instructional effectiveness. Such “gaps” can only become visible when multiple instruments are concurrently calibrated to continuum. The resulting ordering in item positions may reveal that items of one instrument have unique locations on the continuum. The current evidence assessing this idea is very limited. Hopefully, the counterintuitive result—few evidence supporting the idea—motivates future researchers to improve on the study designs, content of the instruments and psychometric methods applied by van der Lans et al. (2019) to more thoroughly study this idea empirically.

References

- André, S., Maulana, R., Helms-Lorenz, M., Telli, S., Chun, S., Fernández-García, C. M., et al. (2020). Student perceptions in measuring teaching behavior across six countries: A multi-group confirmatory factor analysis approach to measurement invariance. *Frontiers in Psychology*, 11, 273.

- Berliner, D. C. (2004). Describing the behavior and documenting the accomplishments of expert teachers. *Bulletin of Science, Technology & Society*, 24(3), 200–212.
- Bijlsma, H. J., Visscher, A. J., Dobbelaer, M. J., & Veldkamp, B. P. (2019). Does smartphone-assisted student feedback affect teachers' teaching quality? *Technology, Pedagogy and Education*, 28(2), 217–236.
- Browne, M. W. (1992). Circumplex models for correlation matrices. *Psychometrika*, 57, 469–497.
- Cannon, M. D., & Witherspoon, R. (2005). Actionable feedback: Unlocking the power of learning and performance improvement. *Academy of Management Perspectives*, 19(2), 120–134.
- Darling-Hammond, L. (2013). *Getting teacher evaluation right: What really matters for effectiveness and improvement*. Teachers College Press.
- De Jong, R., & Westerhof, K. J. (2001). The quality of student ratings of teacher behaviour. *Learning Environments Research*, 4(1), 51–85.
- Doherty, K. M., & Jacobs, S. (2013). *Connect the dots—using evaluations of teacher effectiveness to inform policy and practice*. National Council on Teacher Quality.
- Fauth, B., Göllner, R., Lenske, G., Praetorius, A.-K., & Wagner, W. (2020). Who sees what? Conceptual considerations on the measurement of teaching quality from different perspectives. *Zeitschrift für Pädagogik. Beiheft*, 66(1), 138–155.
- Ferguson, R. F., & Danielson, C. (2014). How framework for teaching and tripod 7Cs evidence distinguish key components of effective teaching. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems*. Wiley.
- Goe, L., Bell, C., & Little, O. (2008). Approaches to evaluating teacher effectiveness: A research synthesis. *National Comprehensive Center for Teacher Quality*.
- Hill, H., Charalambous, C. Y., & Kraft, M. A. (2012). When interrater-reliability is not enough: Teacher observation systems and a case for the generalizability theory. *Educational Researcher*, 41, 561cat1. <https://doi.org/10.3102/0013189X12437203>.
- Howard, G. S., Conway, C. G., & Maxwell, S. E. (1985). Construct validity of measures of college teaching effectiveness. *Journal of Educational Psychology*, 77(2), 187–196.
- Isoré, M. (2009). *Teacher evaluation: Current practices in OECD countries and a literature review*. OECD Education Working Papers, No. 23. OECD Publishing (NJ1).
- Jöreskog, K. G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika*, 43(4), 443–477.
- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254.
- Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research*, 9(3), 231–251.
- Kyriakides, L., Creemers, B. P. M., & Antaniou, P. (2009). Teacher behavior and student outcomes: Suggestions for research on teacher training and professional development. *Teaching and Teacher Education*, 25, 12–23.
- Kyriakides, L., Creemers, B. P., & Panayiotou, A. (2018). Using educational effectiveness research to promote quality of teaching: The contribution of the dynamic model. *ZDM Mathematics Education*, 50(3), 381–393.
- Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective*, (pp. 319–383). The Netherlands, Dordrecht: Springer.
- Martínez, J. F., Schweig, J., & Goldschmidt, P. (2016). Approaches for combining multiple measures of teacher performance: Reliability, validity, and implications for evaluation policy. *Educational Evaluation and Policy Analysis*, 38(4), 738–756.
- Maulana, M., & Helms-Lorenz, R. (2016). Observations and student perceptions of pre-service teachers' teaching behavior quality: Construct representation and predictive quality. *Learning Environments Research*, 19(3), 335–357. <https://doi.org/10.1007/s10984-016-9215-8>.

- Maulana, R., Helms-Lorenz, M., & van de Grift, W. (2015a). Development and evaluation of a survey measuring pre-service teachers' teaching behaviour: A Rasch modelling approach. *School Effectiveness and School Improvement*, 26(2), 169.
- Maulana, R., Helms-Lorenz, M., & van de Grift, W. (2015b). Pupils' perceptions of teaching behaviour: Evaluation of an instrument and importance for academic motivation in Indonesian secondary education. *International Journal of Educational Research*, 69, 98–112.
- Muijs, D., Kyriakides, L., van der Werf, G., Creemers, B., Timperley, H., & Earl, L. (2014). State of the art—teacher effectiveness and professional learning. *School Effectiveness and School Improvement*, 25(2), 231–256.
- Murray, H. G. (1983). Low-inference classroom teaching and student ratings of college teaching effectiveness. *Journal of Educational Psychology*, 75(1), 138–149.
- Nusche, D., Braun, H., Halász, G., & Santiago, P. (2014). *OECD reviews of evaluation and assessment in education: Netherlands 2014*. OECD Reviews of Evaluation and Assessment in Education, OECD Publishing. <https://doi.org/10.1787/9789264211940-en>.
- OECD. (2016). *Netherlands 2016: Foundations for the future*. OECD Publishing. <https://doi.org/10.1787/9789264257658-en>.
- Peterson, K. D. (2000). *Teacher evaluation: A comprehensive guide to new directions and practice*. Corwin Press.
- Praetorius, A. K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction*, 31, 2–12.
- van de Grift, W. J. C. M., Helms-Lorenz, M., & Maulana, R. (2014). Teaching skills of student teachers: Calibration of an evaluation instrument and its value in predicting student academic engagement. *Studies in Educational Evaluation*, 43, 150–159. <https://doi.org/10.1016/j.stueduc.2014.09.003>.
- van de Grift, W. J. C. M., van der Wal, M., & Torenbeek, M. (2011). Ontwikkeling in de pedagogische didactische vaardigheid van leraren in het basisonderwijs [Primary teachers' development of pedagogical didactical skill]. *Pedagogische Studiën*, 88, 416–432.
- van der Lans, R. M. (2018). On the “association between two things”: The case of student surveys and classroom observations of teaching quality. *Educational Assessment, Evaluation and Accountability*, 30(4), 347–366.
- van der Lans, R. M., Maulana, R., Helms-Lorenz, M., Fernández-García, C.-M., Chun, S., Jager, T., Irnidayanti, Y., Inda-Caro, M., Lee, O., Coetzee, T., Fadhilah, N., Jeon, M., & Moorer, P. (2021). Student perceptions of teaching quality in five countries: A Partial Credit Model approach to assess measurement invariance. Manuscript submitted for publication.
- van der Lans, R. M., van de Grift, W. J., & van Veen, K. (2015). Developing a teacher evaluation instrument to provide formative feedback using student ratings of teaching acts. *Educational Measurement: Issues and Practice*, 34(3), 18–27.
- van der Lans, R. M., van de Grift, W. J., & Van Veen, K. (2018). Developing an instrument for teacher feedback: Using the rasch model to explore teachers' development of effective teaching strategies and behaviors. *The Journal of Experimental Education*, 86(2), 247–264.
- van der Lans, R. M., van de Grift, W. J., & van Veen, K. (2019). Same, similar, or something completely different? calibrating student surveys and classroom observations of teaching quality onto a common metric. *Educational Measurement: Issues and Practice*, 38(3), 55–64.
- van der Lans, R. M., van de Grift, W. J., van Veen, K., & Fokkens-Bruinsma, M. (2016). Once is not enough: Establishing reliability criteria for feedback and evaluation decisions based on classroom observations. *Studies in Educational Evaluation*, 50, 88–95.

Rikkert van der Lans is a Postdoctoral Researcher currently working at the Department of Child and Adolescent Psychiatry, LUMC-Curium Leiden (The Netherlands). Previously he worked as a Postdoctoral Researcher at the Department of Teacher Education of the University of Groningen, as a lecturer in methods and statistics at the Department of Educational Sciences (GION) of the University of Groningen, and as a lecturer in the field of psychometrics at the Department of

Methods and Statistics of the University of Tilburg (The Netherlands). His research focuses on the evaluation of professional development and the use of questionnaire and observation data to inform decisions.

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.