Chapter 4

Combining student ratings and classroom observations in teacher evaluation

Abstract

Using item response theory (IRT) this study explores whether items of a student questionnaire and a classroom observation instrument can be viewed as measuring one latent construct, namely teaching skill. The data comprises 269 lessons of 141 teachers which were evaluated using the international comparative analysis of learning and teaching (ICALT) observation instrument and the "My Teacher" student questionnaire. Rasch model analysis confirms that items from both instruments fit a one-dimensional cumulative ordering. Also, students and observers are found to interpret items measuring similar teaching practices equally. However, students and observers can still disagree on which teaching practices any particular teacher uses. After removal of biased items, the correlation between student ratings and classroom observations remains moderate (r = .34).

Based on the manuscript: Van der Lans, R. M., Van de Grift, W., J., C., M., & Van Veen, K. (2016). *Synthesizing teacher evaluation methods: The case of student ratings and classroom observations of teaching*. Manuscript submitted for publication.

4.1 Introduction

This study is motivated by a practical problem: specifically, student ratings and classroom observations provide different evaluations to teachers (e.g., Feldman, 1989; Maulana & Helms-Lorenz, 2016), yet both evaluation methods are considered valid and reliable (e.g., Benton & Cashin, 2012; Kane et al., 2012). This situation considerably complicates the evaluation practice: Although research communicates that schools, districts, and states can choose between these evaluation methods to provide teachers with reliable and valid evaluations, the same research indicate that teachers receiving poor evaluations from one method might have received good evaluations from the other method.

This contradictory situation is partially due to a research tradition in which investigators study the psychometric properties of instruments in isolation of other instruments designed to measure the same latent construct. As a result, these instruments may show high reliability individually, though, when compared with other instruments, show considerable diversity. To reduce this diversity, some researchers advise averaging multiple evaluation methods into a single composite scores (e.g., Mihaly, McCaffrey, Staiger & Lockwood, 2013; Peterson, 2000; NCTQ, 2013). The logic behind this advice is that if both methods measure the same construct, then the average composite score should be considered more reliable and less susceptible to specific biases associated with each particular method. However, because it is unclear whether student questionnaire and classroom observation methods measure the same latent construct, it is unclear whether this composite actually represents one latent variable or constitutes a mix of two distinct constructs.

Researchers put forth various theoretical explanations for the different evaluations between students and classroom observers. Kunter and Baumert (2006) and Maulana and Helms-Lorenz (2016) propose that different evaluation methods will show overlap but that each method also measures specific and unique elements of the construct: teaching skill. These authors maintain that any attempt to synthesize evaluation methods will remove the valuable information that makes each method unique and thus are not worth pursuing. Others question students' expertise to evaluate teaching (e.g., Peterson, 2000), especially younger students and students in lower educational tracks, who may lack the reading and comprehension skills necessary to provide valid item responses. In this line of thinking, classroom observers are trained experts and thus provide more valid indications about teaching than students, who are untrained novices. Further complicating these discussions,

virtually no information about the differences in interpretation between classroom observers and students is available at the item level (see also Maulana & Helms-Lorenz, 2016). Therefore, currently we can only speculate about whether the varying evaluation outcomes reflect differences in interpretation at the item level.

In addition, we note the issue of each method's cost-efficiency. Student questionnaires are more cost-efficient than classroom observation, which makes them an attractive option to replace classroom observation (e.g., Van der Lans, Van de Grift, & Van Veen, 2015; Keuning, Van Geel, Visscher, & Fox, 2016). However, when different methods provide different evaluations to teachers for reasons not completely understood, such decisions may have unintended consequences.

In this study, we use item response theory (IRT) to explore similarities in item responses between students and classroom observers responding to items that measure the same latent variable: teaching skill. Our aim is to explore the factors that underlie the disagreement between students and classroom observers and verify whether they can be attributed to differences in the interpretation of the items and, thus, the construct.

4.2 Background

4.2.1 Context and purpose of the instruments

Current educational policies put increasing emphasis on teacher evaluation and assessment (e.g., Looney, 2011; Mourshed, Chijioke, & Barber, 2010). This study took place within the Netherlands, in which also a growing need has arisen for schools to evaluate and reward effective teaching and to give teachers specific advice on how to improve their teaching (Organisation for Economic Co-operation and Development [OECD], 2016).

The two instruments investigated in this study were developed in this context. The observation method is the international comparative analysis of learning and teaching (ICALT), and the student rating instrument is the "My Teacher" questionnaire. The two instruments have been in use for several years and in various projects, including a national teacher induction project, a regional project directed at improving teacher evaluation at low-performing schools (Van de Grift, 2014), and multiple global projects in which the instrument properties are compared across various countries. Using data from these projects, the reliability and validity of evaluative decisions and feedback based on the two instruments has been investigated thoroughly (e.g., Van de Grift, Helms-Lorenz, & Maulana, 2014, Van der Lans, Van de Grift & Van Veen, 2016, Van de Grift, 2014).

Specific attention has been given to the provision of feedback. To improve teacher feedback, these works connect teaching effectiveness literature with theory on teacher development to arrive at a conceptual understanding how teachers develop skill in and may learn effective teaching practices. This understanding is fundamental to our current aims and is briefly elaborated upon.

4.2.2 The conceptual understanding of how the instruments evaluate teaching skill

An important assumption behind these evaluation policies is that when provided with a reliable identification of a teacher's skill in teaching, evaluators can provide valid and specific feedback regarding what to improve. In this context, if teachers receive no clues about what to improve, their performance evaluation provides little perspective. As Firestone (2014) argues, this situation could even have detrimental effects on education in that it may demotivate teachers. Therefore, it is important that evaluation instruments clarify how they define and conceptualize teaching skill and how they relate current teaching skill to specific advice on how to improve current teaching.

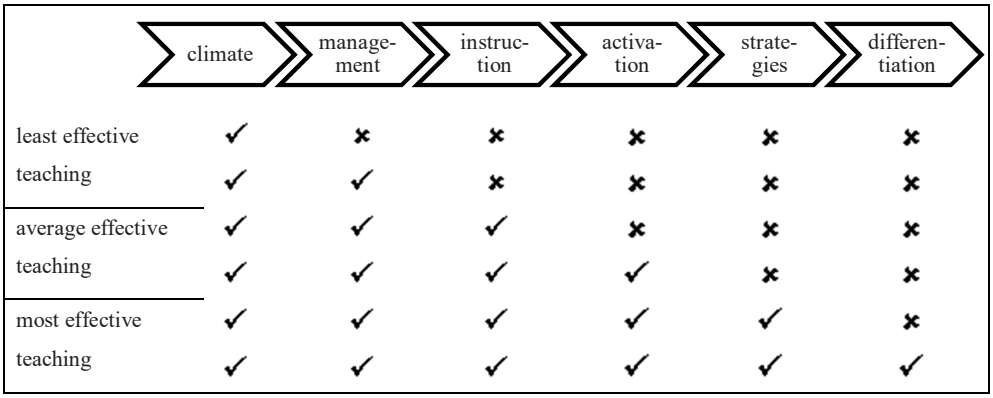
We conceptualize teaching skill as effectiveness. Effective teachers have a large repertoire of teaching methods, behaviors, and strategies (hereinafter, simply "teaching practices") that positively relate to student achievement, such as the practices mentioned in the work of Muijs et al. (2014), Marzano (2003), and Strong (2011). Because the instruments are grounded in literature on teacher effectiveness, the items in the instruments show considerable overlap with items mentioned in other classroom observation and student rating instruments, including the Tripod (e.g., Bill & Melinda Gates Foundation, 2012), classroom assessment scoring system (CLASS) (e.g., Pianta & Hamre, 2009), and the framework for teaching (FFT) (e.g., Danielson, 2013).

To conceptualize improvement, we rely on stages, or cumulative development. In this conceptualization, teachers improve their teaching skills when they successfully add a practice to their repertoire. Moreover, if improvement in teaching skill is conceptualized as cumulative, it logically follows that less complex practices must precede more complex ones and that success in adding a practice will depend on whether its complexity is on par with the teacher's specific skill. In line with Fuller's (1969) three-stage theory of teacher development, we hypothesize a hierarchical and cumulative development in effective teaching practices (for details, see Van der Lans, Van de Grift, & Van Veen [2015, 2016], who discuss how the proposed cumulative ordering fits other work on teacher development,

including Berliner [2001] and Huberman [1993]). Teaching practices involving classroom climate and respectful relationships are the least complex to develop, and competence in them is a prerequisite condition to competence in moderately complex practices such as classroom management and basic instruction. Competence in moderately complex practices, in turn, is a prerequisite to more advanced, complex teaching practices, including interactive instruction, teaching learning strategies, and differentiation. Previous work (Van der Lans, Van de Grift, & Van Veen, 2015, 2016) further refines these three stages into six cumulatively ordered domains (see Figure 4.1) (For a detailed description of the domains, see Van de Grift 2014.)

Figure 4.1

Hypothesized cumulative ordering in domains of teaching practices. Check marks reflect positive observations, and crosses signify negative observations.



Previous results about the cumulative ordering show considerable consistency across these works, but results regarding the ordering of the two most complex domains (i.e., learning strategies and instruction differentiation) is mixed. Specifically, research using the classroom observation method in primary education shows learning strategies to be the most complex (e.g., Van de Grift, Van der Wal & Torenbeek, 2011), a finding corroborated by research using the student questionnaire “My Teacher” in secondary education (e.g., Van der Lans, Van de Grift, & Van Veen, 2015). However, research using the classroom observation method in secondary education shows differentiation to be the most complex (e.g., Van de Grift, Helms-Lorenz, & Maulana, 2014).

4.2.3. In search of complementary evaluations

Cumulative ordering proves valuable feedback to teachers, such that they can understand what they have achieved already, what they should learn (which we refer to as the teacher's "zone of proximal development"), and what is yet too complicated to implement. However, on occasion evaluators find that students give other indications about the teachers' current teaching skill—and thus teachers' zone of proximal development—than observers. The disagreement between classroom observers and students is confirmed by other empirical works. Howard, Conway, and Maxwell (1985) report a correlation of r = .24 between classroom observers' scores for a one-time lesson visit and student ratings. Similarly, De Jong and Westerhof (2001) report an average correlation of r = .12. In recent work using the same classroom observation instrument and student questionnaire used in this study, Maulana and Helms-Lorenz (2016) report that the correlation between scores from classroom observers' one-time lesson visits and student ratings is r = .26. Again, this correlation suggests little overlap, leading the authors to question whether students and observers measure the same construct.

4.3.4 This study

This study applies IRT to investigate in more detail what underlies the disagreement between students' and classroom observers' ratings. Given the low correlations found in previous studies, it seems unrealistic to expect agreement between students and classroom observers on the competency of any particular teacher. However, this disagreement does not necessarily imply that they disagree on their interpretation of the items. Students and observers may order items describing equal teaching practices similarly and assign these similar complexity. In this particular case students and classroom observers agree on the complexity of teaching practices and the low correlation reflects disagreement about which practices the teacher uses. The latter might be explained by the different information that students and classroom observers can access. Students have experienced all the lessons with a teacher and their rating generalize across many lessons, whereas classroom observers have only a snapshot; if teaching skill varies from lesson to lesson, the observer may have an overly positive or negative snapshot and therefore position the teacher accordingly, decreasing the correlation with student data. Thus, the focal research question is as follows: To what extent do observers and students agree on the cumulative ordering in teaching practice complexity?

4.3 Method

4.3.1 Data

Data is selected from three different research projects in the Netherlands. The first is an independent research project focused on the evaluation of in-service teachers working at 13 schools located across the Netherlands. The second is a research project, funded by the Dutch ministry of education, and is located in the Northern provinces in the Netherlands. It focuses on the implementation of teacher evaluation in 11 weak performing schools as judged by the Dutch inspectorate of education. The third project is a ministry financed project focused on evaluation and improvement of beginning teachers (≤ 3 years of experience).

The sample comprises 269 classroom observations of 141teachers having varying experience (0-40 years). The 141 teachers are evaluated by 1,237 students of which 46.3% is male and student age ranged between 11 and 18 years ($Mdn_{(age)}$ = 14 years). All types of education are included including preparatory secondary vocational education, preparatory higher vocational education, and university preparatory education. Class size varied from 5 students (in vocational education) to 30 students (class-size mode is 24 students). The same 141 teachers are also evaluated by 93 observers with varying range in teaching experience (0-40 years). All observers are trained. Inter-rater agreement varied between schools and research projects but all above 70%.

4.3.2 Instruments

"My Teacher" student questionnaire. The "My Teacher" questionnaire has calibrated and validated in two previous works (Maulana et al., 2015; Van der Lans et al., 2015). The study by Maulana had a particular focus on beginning teachers in secondary education. The study by Van der Lans had a particular focus on in-service teachers. The complete questionnaire counts 40 items, some selected by Maulana et al. others by Van der Lans et al. Because the current sample includes in-service teachers having various years of experience, the subsample of 28 items previously identified by Van der Lans, et al. (2015) is used for this study. Each item reflected a statement related to the teacher's teaching practices. Example items are: "my teacher applies clear rules", "my teacher stimulates my thinking", and "my teacher ensures that I use my time effectively." Items can be grouped in six domains: safe learning climate (SLC), efficient classroom management (ECM), quality of instruction (QOI), activating teaching methods (ATM), teaching learning strategies

(TLS), and differentiation in instruction (DII). An extensive review accounting for these six domains is presented by Van de Grift (2014). Students rate items on a dichotomous scale coded "0" = rarely and "1" = often.

The international comparative analysis of learning and teaching (ICALT) observation instrument. A subsample of 31 items of the ICALT observation instrument is used (complete instrument counts 32 items). The selection is again based on previous work which indicates that these 31 items fit the cumulative ordering (Van der Lans, Van de Grift, & Van Veen, 2016). An example item is "uses teaching methods that activate students." Items can be grouped in the same six domains SLC, ECM, QOI, ATM, TLS, and DII. Observers score the items using 4-point scale: 1 = not performed, 2 = insufficiently performed, 3 = sufficiently performed and 4 = well performed. To make comparison possible the, original coding 1 and 2 are recoded 0 and, the original coding 3 and 4 are recoded 1.

4.3.3 Data preparation, cross validation and missing values

In the complete dataset, each classroom observation of teacher "A" is accompanied by an entire class of student ratings. This provides a complex dataset, which is modeled using multilevel Rasch model approach (e.g., de Boeck et al., 2011; Doran, Bates, Blies & Dowling, 2007). However, fit tests for the multilevel Rasch model currently are not implemented in standard statistical software and, therefore, we need to rely on the regular Rasch model to assess model fit. To assess model fit, we randomly selected one student out of each class and connected them with their corresponding classroom observation. This resulted in a dataset comprising of 59 items, i.e. the sum of the number of items comprising the classroom observation instrument (=31) plus the number of items comprising the student questionnaire (=28).

A cross-validation procedure is used to check whether the results are not based on an accidental random selection of students. We randomly selected a second group of students to construct a second sample. This validation sample contained the same classroom observations but in combination with a different set of students.

Development sample. Some classroom observations have missing values on more than one-third of the 31 item responses (n = 10). These are discarded from the analysis. Furthermore, three classroom observations count less than two valid item responses within one of the six domains and are also discarded (n = 3). Using these same criteria all selected

student questionnaires were eligible. After removal of these 13 cases, the sample counted 256 classroom observations corresponding to 256 student ratings. These 256 cases counted 120 missing values, which is 0.8% of all 15,104 item responses.

Validation sample. The same classroom observations are included in the validation sample. Again the 10 observations counting more than one third missing item responses and the three observations counting less than two valid item responses within a domain are discarded. Again all student questionnaires were eligible. The validation sample counted 256 classroom observations connected with 256 student ratings. These 256 cases counted 131 missing values, which is .9% of the 15,104 item responses.

4.3.4 Design and analysis plan

The analysis first concentrates on the ordering of items. Previous research shows that the items of the "My Teacher" questionnaire and the ICALT observation each fit Rasch model assumptions. In this study, it is examined whether they together also fit the Rasch model assumptions. If the Rasch model fits, this would provide evidence that disagreement between observers and students does not reflect that they measure different constructs.

4.3.5 Models and software

As a first step, we analyzed whether the items of different instruments together fit the Rasch model assumptions. The Rasch model specifies three model assumptions. The three assumptions are: local independence, one-dimensionality, and parallel item characteristic curves. Local independence implies that item residuals are uncorrelated. In this study, local independence is assessed using Ponocny's (2001) T_1 and T_{1m} statistics. These are included in the R package eRm (Mair & Hatzinger, 2007). The study examines one-dimensionality as the consistency of an item's complexity (or b-parameters) across random subgroups (Andersen, 1973). To evaluate this consistency, the original sample will be randomly split ten times in two equal halves. Using the Andersen (1973) log-likelihood ratio (LR)-test, we evaluated whether for each random split item complexity in both subgroups remains similar. Finally, parallel item characteristic curves (ICC) evaluate whether item complexity remains similar between varying levels of teaching skill. The Andersen (1973) Log-likelihood ratio test is used again, but now the sample is split using the median teacher evaluation total score.

In the subsequent step, the items that fit the cumulative one-dimensional ordering are selected. The complete sample is used including all student questionnaires. Using the R package lme4 (Bates et al., 2014), a multilevel Rasch model is specified (e.g., de Boeck et al., 2011) to estimate the complexity of the item parameters of the student questionnaires and the classroom observation instrument. We specified a random effect for observer—which could either be a student nested in a class or an observer—, a random effect for teacher and a random effect for item. The R package arm (Gelman et al., 2015) is used to generate the item standard errors.

4.4 Results

Results are reported in three subsequent steps. First Rasch model fit with the three assumptions is evaluated in the development sample. In this step we report about items that misfit the model and which need to be discarded. Then, model fit to the remaining item set is reevaluated in the validation sample. No further item selection is attempted and the results focus is on whether the selected items again fit the Rasch model assumptions. In the third step, the cumulative ordering is reported and discussion focuses on whether the students and classroom observers evaluate practices classified as belonging to the same domain as being approximately similar in complexity. Finally, we turn to the correlation between the student questionnaire and classroom observation method to verify whether resolving measurement bias increases the correlation between methods.

4.4.1 Evaluating model assumptions

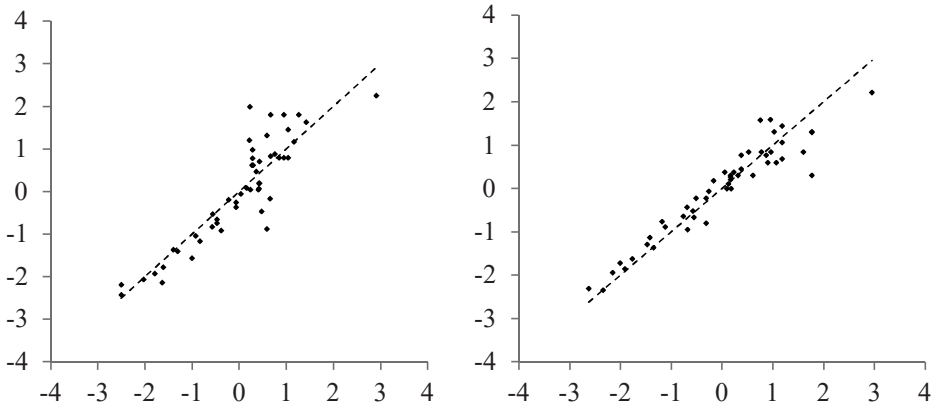
Local independence. Ponocny's T_{1m} statistic diagnoses two "My Teacher" questionnaire items as showing more than one negative residual correlation, in specific: S27 "Teaches me to summarize," and S28 "Explains how I should study something." The negative residual correlations all involve pairings with ICALT items within the domains activating teaching methods (ATM) and teaching learning strategies (TLS). To improve model fit, these two items are discarded. Ponocny's T_1 statistic, identifies 27 positive residual correlations. Two broad patterns are evident. First, residual correlations all involve pairings of items from the same method: i.e., student-student or observer-observer. Secondly, the number of positive residual correlations is greater for the observation instrument and these mostly involve items within the domains: differentiation in instruction (DII) and TLS. After removing 7 items, the remaining 50 items show two decreasing residual correlations and <10 increasing

residual correlations. The list is considered to be sufficiently locally independent. Removing these items did not seem to result in an unacceptable loss of information. Both instruments still cover all six domains.

One-dimensionality. Using a random number algorithm, the sample was ten times randomly split in two. Andersen (1973) Log-likelihood Ratio-tests shows significant deviations. Test values range from $\chi^2(df = 49) = 38.75$, $p = .85$ to $\chi^2(df = 49) = 55.72$, $p = .24$. This suggests that the items have approximately similar cumulative ordering for any random selection of teachers. Using a Goodness-of-Fit (GoF) plot, Figure 4.2 graphically portrays the consistency in item ordering. In a GoF-plot the item ordering of one subsample is plotted against the ordering in the other subsample. The solid line presents the item b-parameters in the first subsample, and the dots represent the b-parameters in the other subsample. The distance of each dot to the solid line indicates the difference in item b-parameters between the two subsamples.

Figuur 4.2

The Goodness of Fit (GoF) plot fo the least (left) and best (right) fitting subgroups.



Parallel ICC. To test the assumption of parallel ICC, the Andersen Log-likelihood Ratio-test (LR-test) (1973) is used. Here, the LR-test examines whether item complexity is approximately similar for teachers evaluated as having above average teaching skill and teachers evaluated as having below average teaching skill. The test included 50 items. Test results suggest that items approximately have parallel ICC ($\chi^2(df = 49) = 66.26$, $p = .051$).

4.4.2 Validation of Rasch model assumptions

The findings in the development sample are reassessed in the validation sample. Ponocny's T_{1m} diagnosed five item pairs violating local independence due to negative residual correlations. Two items; O32 "asks students to reflect on approach strategies" and O17 "boosts the self-confidence of weak students" counted more than one violation. These two items also had been diagnosed in the development sample but these were then considered acceptable. Based on this additional information, we decided to remove these two items and continue with the remaining 48 items. The 48 items counted one negative residual correlation. The T_1 statistic diagnosed 10 item pairs violating local independence due to positive residual correlations.

One-dimensionality – in terms of consistency in item ordering – is not violated. The Andersen LR-test values range from $\chi^2(df = 47) = 29.81$, p = .98 to $\chi^2(df = 47) = 63.36$, p = .06. Also, the Andersen LR-test showed no violations of parallel ICC assumption ($\chi^2(df = 45$[2]$) = 47.29$, $p = .38$. In sum, beside these few violations of local independence this set of items is found to measure an identical construct.

4.4.3 The progressive development in teaching skill

The Table 4.1 shows the established cumulative item ordering. The domains are abbreviated: safe learning climate (SLC), efficient classroom management (ECM), quality of instruction (QOF), activating teaching methods (ATM), teaching learning strategies (TLS), and differentiation in instruction (DII). The Table shows how the predicted ordering evolves in comparable pace among the two instruments and that items in both evaluation methods cover all six domains. This further adds to the validity of the "My Teacher" questionnaire and ICALT observation.

The comparability between classroom observation items and student questionnaire items is sometimes striking. For example, the classroom observer item O4 "ensures mutual respect" ($b = -.77$) and the student questionnaire item S8 "ensures that I treat others with respect" ($b = -.76$) receive almost identical item parameters, suggesting that observers and students agree about the complexity of this aspect of teaching. Also, the list provides information about how students interpret items. For example, the item S40 "helps me if I do not understand" is assigned similar complexity as the item O3 "supports students self-

[2] Items O5, and S24 were excluded from the analysis due to a full response pattern in the more skilled teaching subgroup.

confidence." This result suggests that student responses are triggered by the word "help," (associated with "supports") while less by the word "understand."

Table 4.1

Resulting cumulative item ordering ranging from least complex teaching practices to most complex teaching practices

domain	item	teaching practice	*b*	$SE_{(b}$
SLC	O1	shows respect for students in behavior and language	−2.18	.353
SLC	O2	creates a relaxed atmosphere	−1.40	.266
SLC	S21	treats me with respect.	−1.15	.246
ECM	O7	ensures effective class management	−1.09	.242
SLC	O3	supports student self-confidence	−1.08	.242
SLC	S40	helps me if I do not understand.	−1.08	.242
ECM	S20	prepares his/her lesson well.	−1.05	.238
QOF	O9	explains the subject matter clearly	−1.02	.239
ECM	O5	ensures that the lesson runs smoothly	−.85	.225
SLC	O4	ensures mutual respect	−.77	.219
SLC	S8	ensures that I treat others with respect.	−.76	.219
QOF	O14	gives well-structured lessons	−.74	.219
SLC	S1	ensures that others treat me with respect.	−.68	.214
ECM	O8	uses learning time efficiently	−.64	.212
ECM	S6	answers my questions	−.51	.205
ATM	S23	ensures that I pay attention.	−.44	.202
ECM	S3	makes clear what I need to study for a test.	−.38	.198
ECM	S19	makes clear when I should have finished an assignment.	−.34	.197
QOF	S24	uses clear examples.	−.32	.195
QOF	S13	explains the purpose of the lesson.	−.31	.195
ECM	S39	involves me in the lesson.	−.30	.196
ECM	S26	applies clear rules.	−.25	.192
QOF	O6	checks during processing whether students are carrying out tasks properly	−.20	.193

--- continues next page ---

domain	item	teaching practice	*b*	$SE_{(b}$
QOF	O15	clearly explains teaching tools and tasks	−.17	.193
QOF	O10	gives feedback to students	−.17	.190
ECM	S2	ensures that I use my time effectively.	−.13	.187
QOF	S33	ensures that I know the lesson goals.	−.13	.187
QOF	O11	involves all students in the lesson	−.07	.185
ATM	O13	encourages students to do their best	−.03	.184
ATM	S17	encourages me to think for myself.	.12	.178
ECM	S12	ensures that I keep working.	.23	.175
ATM	O19	asks questions that encourage students to think	.30	.172
ATM	O16	uses teaching methods that activate students	.37	.170
ATM	S30	stimulates my thinking.	.51	.168
ATM	O21	provides interactive instruction	.54	.167
QOF	O12	checks during instruction whether students have understood the subject matter	.57	.166
ATM	O20	has students think out loud	.60	.165
DII	S25	connects to what I am capable of.	.80	.161
DII	S34	checks whether I understood the subject matter.	.83	.161
TLS	O30	encourages students to apply what they have learned	1.00	.158
TLS	O31	encourages students to think critically	1.38	.155
TLS	S16	teaches me to check my own solutions.	1.43	.155
DII	S36	knows what I find difficult.	1.49	.154
DII	O23	checks whether the lesson objectives have been achieved	1.67	.155
TLS	O28	encourages the use of checking activities	1.82	.155
TLS	O29	teaches students to check solutions	1.87	.155
DII	O25	adapts processing of subject matter to student differences	2.30	.157
DII	O26	adapts instruction to relevant student differences	2.41	.158

Regarding the more advanced teaching practices the picture is still somewhat blurred. Because positive residual correlations tend to cluster around more complex teaching practices, the b-parameters of O25, O26, O28, O29, and 030 are biased. Inspection

of two-by-two frequency tables of these items pairs gives the impression that O25 and O26 most plausibly are estimated as more complex than they actually are (i.e. the number in which they both score incorrect is higher than would be expected on the basis of the model), while O28, O29, and O30, are estimated as less complex as they actually are (i.e. the number in which these item pairs score correct is higher than would be expected on the basis of the model). In addition to this, the "My Teacher" questionnaire items describing more complex teaching practices more frequently show model violations and had to be removed. This is true in particular for the items within the domain teaching learning strategies. In this study only one questionnaire item within the domain teaching learning strategies remained: S16 "teaches me to check my own solutions." Currently, this complicates the comparison between the item parameters of the more complex teaching practices.

4.4.4 The correlation between methods

The correlation without item selection (r = .26) is lower than after removal of biased items (r = .34). The correlation without item selection is identical to the correlation recently reported by Maulana and Helms-Lorenz (2016). Thereby this sample reconfirmed their findings. Also, the results suggest that the low correlation is only partially dependent on items measuring different constructs. Most student items and classroom observation items are found to fit the one-dimensional ordering. Also, items measuring the same domains have similar complexity. When the few biased items are deleted the correlation increases with .08.

We have validated that the student questionnaire and classroom observation instrument measure one latent variable. Therefore, it is possible to average them into one composite evaluation score. These 'True' teaching skill estimates (θ_T) are given in Table 4.2. They are more reliable than those of either one method alone and they are directly related to the item complexities mentioned in Table 4.1. For instance, teachers receiving a theta score -.25 were observed to perform most items above item O6, and they would be advised to give attention to items O15, O10, S2, S33, and O11, because the complexity of these practices is close to the current performance level of teacher and class.

Table 4.2

Teachers evaluation scores. Theta (θ) values correspond to item-parameters in Table 4.1

raw score	raw student score[a]	raw observer score[a]	θ[a]	*SE*[b]	$n_{(teachers)}$
20	11	9	-1.02	.411	1
21	16	5	-0.93	.412	1
22	15	6	-1.00	.402	1
25	12	13	-.80	.411	4
26	13	14	-.73	.411	1
27	9	18	-.72	.405	1
28	11	17	-.70	.408	2
29	16	13	-.55	.411	7
30	15	15	-.55	.403	4
31	14	17	-.47	.407	4
32	17	15	-.36	.412	5
33	19	14	-.25	.403	4
34	17	17	-.25	.414	5
35	18	17	-.18	.413	4
36	18	18	-.14	.410	10
37	19	18	.00	.414	11
38	19	19	.05	.406	11
39	19	20	.10	.413	8
40	18	22	.15	.417	13
41	19	22	.23	.408	15
42	20	23	.40	.420	7
43	20	23	.45	.414	11
44	18	26	.41	.428	1
45	21	24	.71	.416	5
46	20	26	.73	.431	3
47	21	26	1.01	.440	2

[a]. If multiple teachers had similar raw scores, the reported value is the mean.

[b]. If multiple teachers had similar raw scores, the reported value is the median.

4.5 Conclusion and Discussion

This study combined a student questionnaire (the "My Teacher" questionnaire) and observation instrument (the ICALT observation) and explored whether items of both instruments measure one latent variable, namely teaching skill. The general conclusion of this study is that students and observers agree on the complexity of similar teaching practices. This finding is inconsistent with previous speculations that student questionnaire and classroom observation methods must measure different constructs because they offer different perspectives. Clearly classroom observations and student questionnaires *may* result in different measurement. Questionnaires can address aspects of teaching which observers cannot readily observe (e.g., whether students understood the explanation). Also, classroom observation can evaluate aspects of teaching skill that students cannot reasonably evaluate (e.g., the quality of the lesson content and materials). But our results suggest that when observers and students evaluate aspects of teaching they both can observe, their ratings are psychometrically similar and one-dimensional.

This implies that the low correlation between classroom observation and student questionnaires cannot reasonably be explained as being due to both instruments measuring different constructs. While our results replicate the low correlation between students and observers, we also find that student questionnaire items and classroom observation items fit the same one-dimensional cumulative ordering. The correlation between evaluation methods is $r = .26$ and after removal of those few misfitting items increases only slightly to $r = .34$. This slight increase gives reason to doubt whether student questionnaires and classroom observation instruments when they would show perfect one-dimensional measurement—and measure exactly the same construct—would approach a correlation of $r = 1.00$. We need to consider other explanations as to why students and observers disagree on teachers' teaching skill.

4.5.1 Alternative explanations of the low correlation between students and observers

What alternative explanations can be given for the unexpected low correlation? One important explanation may be the low reliability of one-time lesson visits, which provide an unrepresentative picture of the teachers' teaching skill (Kane ,et al. 2012; Van der Lans, et al. 2016). On the other hand, Benton and Kashin (2012) and Marsh (2007) state that student questionnaires are more valid because students have observed all lessons and that they are with many. It might be unreasonable to expect that evaluation outcomes based on a one-

time lesson visit by a single observer should be comparable with evaluation outcomes based on all lessons and observed by many observers. An example is Murray's (1983) study which correlated classroom observations of multiple lessons by multiple observers with student questionnaire data and reports a correlation of $r = .76$. This result suggests that the moderate correlation might be due to low reliability. This suggests that increasing the number of observers and the number of lessons observed could increase the correlation between classroom observation data and student questionnaire data.

4.5.2 Limitations

The study has some limitations which should be taken into account. One limitation involves the violations of local independence. Though these are few, they tend to concentrate around items within more complex domains. This is especially true for the observation instrument. Therefore, the teaching practices in the ICALT domains teaching learning strategies and differentiation in instruction are biased. It may be argued that due to accepting some violations of local independence the correlation is somewhat biased (either too optimistic or pessimistic). Another limitation concerns the cross-validation analysis which contained the same classroom observations twice and only varied the student ratings. It may be argued that the positive cross-validation result is due to using part of the data twice. Finally, a multilevel approach should be preferred when testing clustered data as students nested in classes. We note that the item parameters in Table 4.1 are estimated using multilevel techniques, but acknowledge that the Rasch model fit tests are not corrected for the clustered data structure. Methods to evaluate Rasch model assumptions within a multilevel framework are still in a developmental phase (e.g., de Boeck, et al. 2011). We consider our choice to randomly sample one student from each class as the best option currently available.