Same, Similar, or Something Completely Different? Calibrating Student Surveys and Classroom Observations of Teaching Quality Onto a Common Metric

van der Lans, Rikkert M.; van de Grift, Wim J. C. M.; van Veen, Klaas

Published in:
Educational Measurement: Issues and Practice

DOI:
10.1111/emip.12267

Publication date:
2019

Same, similar, or something completely different? Calibrating student surveys and classroom observations of teaching quality onto a common metric

Abstract

Using item response theory, this study explores whether it is possible to calibrate items contained in a student survey with a classroom observation instrument onto a common metric of teaching quality. The data comprise 269 lessons and 141 teachers, evaluated using the international comparative analysis of learning and teaching (ICALT) observation instrument and the My Teacher student survey. Using Rasch model concurrent calibration, the authors calibrate items from both instruments onto a common one-dimensional metric of teaching quality. Challenges pertain mainly to items measuring teaching students learning strategies and differentiation. The authors detail some explanations for these difficulties.

Worldwide, education initiatives seek to improve teacher evaluation methods, with the goals of enhancing instruction quality and on the job performance (e.g., Isoré, 2009; Doherty & Jacobs, 2013). Teacher performance evaluation holds a strong policy appeal as it focusses on key determinants of educational quality, such as instruction quality, classroom management, and pedagogy. Conventional wisdom indicates that a valid evaluation requires a combination of various measures. The combination of measures arguably should yield a more complete, reliable, and accurate assessment of teacher performance (Goe & Croft, 2009; Kane & Staiger, 2012; Steele, Hamilton, & Stecher, 2010); provide more detailed feedback to teachers (Baker et al., 2010); and increase the cost effectiveness of evaluation efforts (Van der Lans, Van de Grift & Van Veen, 2015; Downer, Stuhlman, Schweig, Martínez & Ruzek, 2015). Even with the recognition of these advantages though, no consensus exists regarding how to combine the measures to achieve these diverse benefits (Martínez, Schweig, & Goldschmidt, 2016). For example, Kane and Staiger (2012) proposed to use composite measures (i.e. the average of multiple measures), because composites yield more reliable evaluations. However, because composite measures tend to be complex to interpret, we believe they offer limited potential to provide teachers with more detailed and meaningful feedback.

We propose that optimal combinations would balance the strengths of some measures against the weaknesses of others, such that they are complementary. For example, classroom observation measures can provide virtually immediate feedback and coaching after a lesson (e.g., Downey, Steffy, English, Frase & Poston, 2004). But gathering multiple observations is cost intensive and single observations suffer from low reliability (Hill, Charalambous, & Kraft, 2012; Praetorius, Pauli, Reusser, Rakoczy, & Klieme, 2014). Student surveys provide high reliability (e.g., Van der Lans, 2018; Marsh, 2007) and are relatively cost efficient, but students cannot provide ongoing coaching or training for teachers. Therefore, optimal teacher

performance evaluations might combine reliable student surveys with feedback derived from classroom observations, such as by giving the observers the results of the surveys, so that they can focus on specific teaching practices (e.g., those that earned them poor scores from students) while observing the lesson.

One condition for such a combined strategy to function, is that the student survey and classroom observation measures can be standardized onto a common metric. This metric in turn needs to have the capacity to link the teacher's performance score back to specific teaching practices in need for improvement. Previous studies have established a means to order the observations of teaching practices according to a one-dimensional scale that features five or six stages of the development of effective teaching (Van de Grift, Van de Wal & Torenbeek, 2011; Van de Grift, Maulana, & Helms-Lorenz, 2014; Van der Lans et al., 2015, 2018; Kyriakides, Creemers, & Panayiotou, 2018). The established stage-order overlaps with those reported in research into teacher development (Berliner, 2004; Fuller, 1969; Huberman, 1993) and provide a means to link the teacher's performance score back to specific teaching practices associated with that stage of development. It has been shown that using these stage models to scaffold feedback and coaching has medium to large effects on development of teaching quality (Tas, Houtveen, Van de Grift & Willemsen, 2018) and may outperform feedback and coaching methods not based on the stages (Antoniou & Kyriakides, 2011). Furthermore, both student surveys and classroom observations have been proven valid measures to identify teachers' stage of development (Van de Grift et al., 2014; Van der Lans et al., 2015, 2017, 2018; Maulana & Helms-Lorenz, 2016). Yet no existing evidence specifies whether the more reliable identification of the teacher's stage obtained from student surveys also can inform the less reliable classroom observations, as they pertain to which teaching practices require coaching and training.

To address this gap, we apply a Rasch model based concurrent calibration approach to determine if classroom observations and student surveys can be calibrated on the same one-dimensional measurement scale with six stages of teaching quality. Our focus is on teachers' instructional practice, even though some student surveys have a much wider scope, including measures of student engagement and attitudes. Although these constructs might inform the larger teaching quality construct, we purposefully focus on observable elements of teachers' instruction efforts. The Rasch model based concurrent calibration approach also offers an alternative means to assess the level of (dis)similarity in measurements (Kolen & Brennan, 2014), in that it seeks to calibrate items from different instruments on the same dimension (or measurement scale). That is, with concurrent calibration, we can test whether classroom observation and student survey items, developed to measure the same six stages, locate items describing similar teaching practices on more or less the same position in the one-dimensional stage ordering. Our primary research question is as follows: *To what extent do student survey and classroom observation items of teaching quality lead to the same operationalization of a one-dimensional measure of teaching quality?*

Background

Instruments

Two instruments are central in this study: the International Comparative Analysis of Learning and Teaching (ICALT) observation instrument and the My Teacher Questionnaire (MTQ). Both instruments aim to measure the same latent construct, teaching quality (Van de Grift et al., 2014; Van der Lans et al., 2015). Teaching quality comprises six latent domains that can be ordered on a single measurement scale (see Figure 1). We briefly describe the six domains; Table 1 provides example items from the ICALT and MTQ related to each domain.

Safe learning climate. The critical role of respectful relationships is corroborated by psychological theory, including attachment (Bowlby, 1969) and self-determination (Ryan &

Deci, 2000) theories. Attachment theory postulates that a safe environment stimulates children to take initiative and explore, because they know that an adult will be there to help them (Bowlby, 1969). According to Pianta and colleagues, the principles of attachment theory generalize to the classroom setting (Hamre et al., 2013) asserting that students who view their teacher as fair and supportive are more likely to discover new things and more likely to actively participate in academic activities (Wentzel, 2002). Also, self-determination theory assigns a key role to respectful relationships in facilitating student motivation and performance (Ryan & Deci, 2000).

Efficient classroom management. Successful classroom management establishes procedures, routines, and rules about where and how learning takes place, as is necessary for instructional activities to be executed successfully (Korpershoek, Harms, de Boer, Van Kuijk, & Doolaard, 2016; Muijs & Reynolds, 2003).

Clear and structured explanation. Clear explanations help students recall their prior knowledge, expand their critical knowledge, and confirm their comprehension of the content (Muijs & Reynolds, 2003; Rosenshine, 1995). Relevant teaching practices stimulate students to engage in cognitive processing of the lesson content. According to Bloom, Engelhart, Furst, Hill, and Krathwohl's (1956) taxonomy, clear, structured explanations help students remember and comprehend facts and procedures.

Activating teaching methods. Activating teaching methods evoke interactions between the teacher and students and among students by requiring that students engage in collaborative group work, explain topics to one another, or think aloud (Abrami, Bernard, Borokhovski, Waddington, Wade, & Persson, 2015; Muijs & Reynolds, 2003). In Bloom et al.'s (1956) taxonomy, activating teaching methods stimulate students to apply and analyze learned material.

Teach learning strategies. When they teach learning strategies, teachers stimulate the development of students' metacognitive skills and self-regulated learning, such as by asking them to explain how they solved a problem or if there might be multiple ways to answer a question (Abrami et al., 2015). In Bloom et al.'s (1956) taxonomy, teaching learning strategies stimulates students to synthesize and evaluate the learned material.

Differentiation in instruction. Teachers should adjust their instructional practice to specific students' learning needs, perhaps by allowing flexible time to complete assignments or providing additional explanations in small groups (e.g., Reis, McCoach, Little, Muller, & Kaniskan, 2011). In terms of Bloom et al.'s (1956) taxonomy, differentiation involves helping low-ability students to remember and comprehend, assisting moderate-ability students to apply and analyze material, and stimulates high-ability students to synthesize and evaluate material.

---------- INSERT TABLE 1 ABOUT HERE ----------

One-dimensional stage-order model. The process of becoming an expert teacher appears to move along specific and sequentially or cumulatively ordered phases (e.g., Berliner, 2004; Fuller, 1969; Huberman, 1993). In consistent findings, Fuller (1969) and Huberman (1993) identify skills for acquiring and maintaining respectful relationships with students as the first phase of teacher development. Berliner (2004) and Fuller (1969) maintain that classroom management and basic instruction routines are prerequisites for more student-centered teaching approaches. Such descriptions relate closely to the six domains, revealing how teaching quality develops (Van der Grift et al., 2014; Van der Lans et al., 2017, 2018), as summarized in the stage-order framework in Figure 1. Kyriakides et al. (2018) report a similar one-dimensional stage-order model with five stages. That is, they also find two initial stages related to classroom management and structuring explanations, and their final two stages are

related to differentiation and modeling (e.g., teaching students self-regulated learning strategies).

---------- INCLUDE FIGURE 1 APPROXIMATELY HERE ----------

Concurrent Calibration of Observation and Survey Measures

Prior studies examining the overlap between survey and observation measures typically apply correlational techniques (e.g., Van der Lans, 2018; Downer et al., 2015; Ferguson & Danielson, 2014; Howard, Conway & Maxwell, 1985; Kane & Staiger, 2012; Martínez et al., 2016; Maulana & Helms-Lorenz, 2016; Murray, 1983; Polikoff, 2015) and report Pearson correlations and uncover modestly sized associations (e.g., 0.20–0.30) between survey and classroom observation total scores. Studies that further decomposed the construct teaching quality into smaller factors have reported associations of similar size. For example, Ferguson and Danielson (2014) correlate the seven subscales of the Tripod survey (caring, controlling, clarifying, challenging, captivating, conferring, and consolidating) with the four subscales of the Framework for Teaching (FFT) (planning and preparation, classroom environment, instruction, professional responsibilities) and find correlations ranging from 0.088 to 0.331. Other studies rely on (multilevel) regressions that allow for the inclusion of covariates, but associations remain of modest size (Downer et al., 2015; Martínez et al., 2016; Polikoff, 2015). These correlational studies show that students and observers score the same teachers different, yet it remains unclear what exactly the students and observers disagree about. They might disagree about the measured construct, about the teachers' skill level, or both.

With Rasch model concurrent calibration, we make strong assumptions about each respondent's item response pattern and the validity of these assumptions can tested independently of the (reliability of) the total score (Bond & Fox, 2007; Rasch, 1960). We present the basic idea in Figure 2: Individual students and observers may exhibit remarkable

consistency about the one-dimensional stage-order, even if they disagree about the exact stage to which to assign teacher A. With a concurrent calibration, we can determine if an individual observer and an individual student who provide similar assessments of the teacher's teaching quality also exhibit an equal probability of endorsing items related to the six domains (e.g., student 5 and observer 1 in Figure 2). We believe this approach can provide novel insights concerning how best to combine student survey and classroom observation items.

---------- INSERT FIGURE 2 APPROXIMATELY HERE ----------

Hypotheses

The similarity of the cumulative ordering of the MTQ and ICALT instruments can be established if items that target the same latent domain appear in similar locations in the stage ordering, as illustrated by teachers A–F in Figure 3.

---------- INSERT FIGURE 3 APPROXIMATELY HERE ----------

Teacher G instead provides an example of an item response pattern in which the stage ordering differs between the MTQ and the ICALT. It implies either a misfit with the cumulative ordering or, if it is a dominant pattern, a fit with the cumulative ordering that is rearranged, such that survey items and classroom observation items each cluster together. We consider two testable hypotheses about the plausibility of the pattern of teacher G:

$H1_0$: The items of either the survey or classroom observation instrument (predominantly) misfit the model.

$H1_A$: The items of both measures (predominantly) fit the model.

$H2_0$: Item position in the cumulative ordering is dependent on the instrument.

$H2_A$: Item position in the cumulative ordering is independent of the instrument.

Method

Data

We selected data 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 low-performing schools as judged by the Dutch inspectorate of education. The third project is also a ministry-financed project focused on evaluation and improvement of beginning teachers (≤ 3 years' experience).[1]

The procedures for the projects varied. In all of them student surveys and classroom observations were spaced apart in time. The two Ministry-financed projects collected data in fall (October–December) and spring (March–May), using a single survey and one classroom observation. In these studies, a single classroom observer might visit the same teacher twice, in which cases we included only one of the observations in this study. The independent research project collected data throughout the school year, though concentrated in January–May. It also gathered up to three observations by three different observers and one survey in the same classroom setting.

The total sample comprises 269 classroom observations of 141 teachers with varying levels of experience (0–40 years). The 141 teachers were rated by 93 observers, who also varied in their teaching experience (0–40 years). All observers were trained. The interrater agreement varies across schools and research projects, but all exceed 70%. The student ratings came from 1,237 participants (46.3% male, 11 to 18 years, median age 14 years), representing all levels of education: (lower) preparatory secondary vocational education, preparatory higher vocational education, and university preparatory education. Class sizes ranged from 5 (in lower vocational education) to 30 students (mode = 24 students).

Measures

[1] Project title "landelijk onderzoek naar inductie effecten van inductie." project number: OCWOND/OD8-2013/45916 U.

From the 40-item MTQ, we selected a subsample of 28 items, in line with previous work that confirms these items fit the hypothesized one-dimensional measure (Van der Lans et al., 2015). Each item contained a statement about the teacher's teaching practices and used a dichotomous rating scale, with 0 = "rarely" and 1 = "often." From the 32-item ICALT observation instrument, we took 31 items, which previous work has indicated provide good fit with the one-dimensional measure (Van der Lans et al., 2018). The classroom observers scored these items on a four-point scale: 1 = "not performed," 2 = "insufficiently performed," 3 = "sufficiently performed," and 4 = "well performed." To support comparisons, we recoded codes 1 and 2 to equal 0 and codes 3 and 4 to equal 1. With a dichotomous Rasch model and polytomous partial credit model, we can estimate the potential effect of the dichotomization. The correlation between the person parameters is r = .92 (df = 246), and the range of person scores is only slightly higher for the dichotomous categories (Min = -2.34; Max = 4.18) than the polytomous categories (Min = -1.09; Max = 4.95). This evidence indicates no substantial differences.

Model and design

The first hypothesis requires testing Rasch model assumptions. This is done in a one-observer-one-student design. In this design each classroom observation is matched with one randomly selected student survey related to the same teacher. Two datasets with this design were produced. The first is referred to as the development sample. The second which matched another randomly selected another student with the classroom observations, is referred to as the validation sample. We can justify these random selections of single students, because we test students' item response patterns independently from the reliability of the total score (Figure 2).

The second hypothesis is tested using a multilevel Rasch model. For this, we organize the data in a long format, listing all ratings by students and observers in one column (De

Boeck et al., 2011). This model includes six facets: item (i), domain (d), method (m) (1 = ICALT, 2 = MTQ), observer (o) (student or classroom observer), teacher (t), and class (c). In g-theory language, the design is as follows: {[(o × i) : m] × d} × (t : c), where × indicates facets that are crossed, : indicates facets that are nested, and the brackets define the reading order. Thus, for example, observer and item are crossed within each method and within each method item; observer and domain also are crossed. This g-study design distinguishes 19 random effects, though the random effect for observer × item × teacher must be confounded with the observer × item facet to ensure model convergence. Accordingly, in Appendix A, we list all 18 random effects accounted for by the models.

Data preparation and missing values

To assess the representativeness of the complete sample, we estimated the correlation of the aggregated student survey total scores with the classroom observation scores. The resulting correlation of $r = 0.26$ is similar to the values reported by Maulana and Helms-Lorenz (2016) and Howard et al. (1985). It signals the sample's representativeness.

Development sample. We excluded classroom observations for which more than one-third of the 31 item responses were missing values (n = 10) and those that had fewer than 2 valid item responses on one of the six domains (n = 3). All the student surveys were eligible though. After removing the 13 observations, the sample consisted of 256 classroom observations, corresponding to 256 student surveys. The cases featured 120 missing responses, or .8% of all 15,104 item responses.

Validation sample. We again excluded 13 classroom observations from the validation sample, and again, all the student surveys were eligible. The validation sample thus included 256 classroom observations connected with 256 other student ratings. These 256 cases featured 131 missing responses, equivalent to .9% of the 15,104 item responses.

Analysis plan and statistical software

To examine the first hypothesis, we test whether the 59 items from the different instruments meet three Rasch model assumptions: local independence, one-dimensionality, and parallel item characteristic curves (ICCs). To assess local independence, we use Ponocny's (2001) T_1 and T_{1m} statistics, included in the R package eRm (Mair & Hatzinger, 2007). We test for one-dimensionality with the consistency in the item *b*-parameters across random subgroups, such that we randomly split the original sample 10 times into two equal halves and examined whether the *b*-parameters in both subgroups remain similar, according to Andersen's (1973) log-likelihood ratio test (LR test). Finally, the Andersen (1973) LR-test also evaluates parallel ICCs, but instead of a random split, it splits the sample according to the median teacher evaluation total score.

To test the second hypothesis that predicts items' positions on the measurement scale depend on the instrument, we use the R package lme4 (Bates, Maechler, Bolker, & Walker, 2014). When we visually inspected the item parameters for the MTQ and ICALT using a multilevel Rasch model, we could identify random effects for class, observer (which could be a student nested in a class or an observer), teacher, and item. To estimate the standard errors, we used the R package arm (Gelman et al., 2015). Finally, with a chi-square difference test, we determined if excluding the random effect method×domain decreases mode fit.

Results

In this section, we designate classroom observation items with an O (e.g., O2 and O5 refer to classroom observation items 2 and 5) and student survey items with an S (e.g., S7 and S20 indicate student survey items 7 and 20).

Hypothesis 1: Evaluating Rasch model fit in the development sample

Local independence. Ponocny's (2001) T_{1m} statistic identifies two "My Teacher" survey items that indicate more than one negative residual correlation: S27, "Teaches me to summarize," and S28, "Explains how I should study something." The negative residual

correlations all involve pairings with ICALT items in the activating teaching methods and teaching learning strategies domains. To improve model fit, we discarded these two items. The T_1 statistic also identifies 27 positive residual correlations. Two broad patterns emerge. First, residual correlations all involve pairings of items from the same method (i.e., student–student or observer–observer). Second, the number of positive residual correlations is greater for the observation instrument, and they mostly involve items pertaining to the differentiation in instruction and teaching learning strategies domains. After we removed 7 items, the remaining 50 items indicate two decreasing residual correlations and fewer than 10 increasing residual correlations. We considered the list sufficiently locally independent. Moreover, removing these items does not result in an unacceptable loss of information since both instruments still cover all six domains.

One-dimensionality. With a random number algorithm, we split the sample 10 times. Andersen's (1973) LR-test values range from $\chi^2(df = 49) = 38.75$, $p = .85$, to $\chi^2(df = 49) = 55.72$, $p = .24$, which suggest that the items display approximately similar cumulative ordering for any random selection of teachers. Using a goodness-of-fit (GoF) plot, Figure 3 graphically portrays the consistency in item ordering. In a GoF plot, the item ordering of one subsample gets plotted against the ordering in the other subsample. The solid line represents the item *b*-parameters in the first subsample; the dots represent the b-parameters in the other subsample. The distance of each dot to the solid line indicates the difference in the items' *b*-parameters between the two subsamples.

---------- PLEASE INSERT FIGURE 4 APPRXIMATELY HERE ----------

Parallel ICCs. To test the assumption of parallel ICCs, we use Andersen's (1973) LR-test and examine whether item complexity is approximately similar for teachers evaluated as having above-average or below-average teaching skill. The test, which includes 50 items, suggests the items have approximately parallel ICCs ($\chi^2(df = 49) = 66.26$, $p = .051$).

Hypothesis 1: Reconfirming Rasch model fit in the validation sample

We reassessed the development sample findings with the validation sample. Ponocny's (2001) T_{1m} test diagnosed five item pairs that violate 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 were also identified in the development sample but appeared acceptable in that case. With this additional information, we decided to remove these two items and continue with the remaining 48 items, which had one negative residual correlation in the validation sample. The T_1 statistic diagnosed 10 item pairs that violated local independence due to positive residual correlations which we considered acceptable.

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 $\chi^2(df = 47) = 63.36$, $p = .06$. In addition, this test showed no violations of the parallel ICC assumption ($\chi^2(df = 45) = 47.29$, $p = .38$).[2] Therefore, except for a few violations of local independence, this set of items broadly fits the one-dimensional cumulative ordering.

Hypothesis 2: Differences in item position between instruments

To evaluate the second hypothesis, we assessed whether the variability in *b*-parameters depends on the method after we account for their dependency on the domain. We estimate two nested multilevel Rasch models, one without the domain × method interaction and another that includes this facet. The chi-square difference test is significant ($\Delta\chi^2$ ($df = 1$) = 4.10, $p = 0.043$), indicating an absolute difference in the *b*-parameters between survey and observation items related to the same domain. Further inspection reveals that difference in *b*-parameters is almost completely due to the domain differentiation. If we remove items related to this domain, adding the domain × method interaction is no longer predictive ($\Delta\chi^2$ ($df = 1$) =

[2] We excluded items O5 and S24 from the analysis because of the full response pattern in the more skilled teacher subgroup.

0.0, $p > 0.05$). Thus, the selected subset of student survey and classroom observation items' *b*-parameters are independent of the method, except for items related to differentiation.

The combined measurement scale

Table 1 contains the established cumulative item ordering of the instruments combined. The ordering is estimated using the multilevel Rasch model design.

---------- PLEASE INSERT TABLE 2 APPROXIMATELY HERE ----------

The comparability between classroom observation items and student survey items is sometimes striking. For example, item O11 ("involves all students in the lesson"; $b = .03$) and item S39 ("Involves me in the lesson"; $b = .06$) receive almost identical *b*-parameters, suggesting that observers and students agree about the complexity of this aspect of teaching. The comparability between items O8, "uses learning time efficiently" ($b = -.56$), and S2, "ensures that I use my time effectively" ($b = -.10$), also is notably large. In this sense, Table 1 is informative about differences in item difficulty, but it provides only a visual indication of whether the *b*-parameters depend on the instrument.

Discussion

In response to our research question, we uncover tentative support for the effort to calibrate student survey items and classroom observation items on a common measurement scale; it appears possible to calibrate these items on the same scale, though perhaps not for all domains of effective teaching. The specific results indicate few problems with items in domains associated with a safe learning climate, efficient classroom management, clear and structured explanations, and activating teaching methods. However, the results for teaching learning strategies and differentiation in instructions are mixed, and our further exploration suggests that the challenges for calibrating items in these two domains are distinct.

For differentiation, we encounter no significant problems when calibrating the items to the cumulative measurement scale. Six of the seven items fit, including three observation and three student survey items. Thus, we retain $H1_A$ for differentiation: Items predominantly fit the measurement scale. However, we also determine that item position depends on the instrument, as is even evident in Table 1, because all three student survey items exhibit lower *b*-parameters than the items from the classroom observation instrument. Thus, we reject $H2_A$ for this domain: Item position in the cumulative ordering is not independent of the instrument.

With respect to learning strategies, we faced significant challenges to fit the items to measurement scale. Of the nine items, only five fit: four from the observation instrument and one from the student survey. Even though items from both instruments fit, the number of survey items is at the absolute minimum. Thus, with regard to the learning strategies domain, we reject $H1_A$, in that items do not predominantly fit the measurement scale. The second analysis instead shows no significant dependence on the instrument. Thus, the item positions are approximately similar, so in this case, we confirm $H2_A$, and conclude that item position is independent of the instrument.

Potential explanations of encountered problems: differentiation in instruction domain

To explain the observed differences in item position, we seek potential factors that do not affect model fit but can differentially influence item difficulty (*b*-parameters) across instruments. Potential explanations consistent with these findings may relate to student characteristics, such as age and maturity; observer characteristics; or differences in item phrasing. We consider two possible explanations.

Observer characteristics. Scriven (1981) claims that (trained) classroom observers' scorings reflect common standards and norms about teaching. In the Netherlands, various policy agents have called attention to the complexity and difficulty of adapting instruction to individual student needs (e.g., Dutch Inspectorate of Education, 2016). This call has had a

profound impact, prompting a widely shared consensus among teachers, school leaders, and researchers about the challenges of differentiated teaching. Such consensus in turn may have biased classroom observers to overrate the complexity of adapting their explanations. The only available evidence for this explanation is the greater number of violations of local independence among the classroom observation items associated with the differentiation domain (see also Van der Lans et al., 2018). These violations do not arise among the student survey items and thus seem unrelated to the measurement of the domain in general. The violations suggest that observers score the items associated with adapting explanations more similarly than would be expected by the model, consistent with the notion that social consensus or norms might influence the scoring of classroom observation items related to a differentiation domain.

Item phrasing. Another explanation might relate to the item content in the "My Teacher" student survey. It is debatable whether items such as "connects to what I am capable of" provide a similar operationalization of the differentiation domain, relative to classroom observation items such as "adapts processing of subject matter to student differences" or "adapts instruction to relevant student differences." Notably, the survey items appear less specific to the instructional situation, without detailing whether the teacher acknowledges student capabilities by explaining the same assignment or material with varying complexity or at a different pace (adaptation of processing) or by giving the student different assignments or materials (adaptation of instruction). The survey item "connects to what I am capable of" even may refer to both situations, such that it might be scored more positively. The classroom observation instrument is more specific about such instructional differences, though the larger number of positive residual correlations suggests that observers experience difficulties distinguishing between these instructional elements.

Potential explanations for encountered problems: learning strategies domain

To explain model misfit, we seek potential factors that do not affect the item position but that lead to inconsistencies in the item ordering. Perhaps student age may cause misfit of the item response pattern. If young students misunderstand the item content, it could lead to random-like item response pattern that misfit the model. However, such an outcome likely would also affect item positions and, thus lead to the rejection of $H2_A$. Another explanation holds that some but not all students have had any experience with teachers performing learning strategies. The teaching practices associated with the learning strategies domain are complex and practiced by relatively few teachers. Hence, perhaps students having no experience with teachers applying learning strategies have different understanding of the item content compared to the students having experience with teachers that applied learning strategies.

Limitations

The study's conclusions are restricted by the specific instruments used. The sample is limited in size. Therefore, the results should be interpreted with caution and should encourage further research that uses concurrent calibration methods. The cross-validation analysis only varied the student ratings; the positive cross-validation result thus could arguably result from using the classroom observation data twice.

Potential practical implications and directions for future research

Evaluating teacher performance through observation is complex and expensive; complementing classroom observations with student survey measures potentially offers the promise of correcting the "snapshot" provided by observations, by providing a more general image, derived from students' perspectives of teachers' lessons. In the introduction we proposed to use student surveys to inform classroom observers and coaches about teachers' stage of development. This way schools and districts can better target their classroom observation and coaching efforts. One condition of success for this approach is that the

student survey and classroom observation items can be standardized to the same metric. The study results suggest that this condition can be met for four or perhaps even five out of the six measured domains. Yet, other questions remain. An important one pertains to the nature of the unreliability in classroom observations and student surveys. The high reliability of student surveys is mainly due to the sampling of raters (Marsh, 2007), whereas the reliability of classroom observations is among other facets dependent on the number of sampled lessons (e.g., Praetorius et al., 2014). Thus, in practice it may turn out that teaching practices identified by the students as in need for improvement are not part of the specific lesson (occasion) visited by the classroom observer. Hence, further research is needed to verify whether the proposed procedure truly enhances the cost effectiveness of feedback and coaching.

References

Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A., & Persson, T. (2015). Strategies for teaching students to think critically: A meta-analysis. *Review of Educational Research*, *85*(2), 275-314. doi: 10.3102/0034654314551063

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38*(1), 123–140.

Antoniou, P., & Kyriakides, L. (2011). The impact of a dynamic approach to professional development on teacher instruction and student learning: Results from an experimental study. *School Effectiveness and School Improvement*, *22*(3), 291-311.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., . . . Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers*. Washington, DC: Economic Policy Institute.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using S4 classes*. R package version 1.1-7. URL: http://CRAN.R-project.org/package=lme4

Berliner, D. C. (2004). Expert teachers: Their characteristics, development and accomplishments. In R. Batllori i Obiols, A. E Gomez Martinez, M. Oller i Freixa, & J. Pages i Blanch (Eds.), *De la teoria….a l'aula: Formacio del professorat ensenyament de las ciències socials* (pp. 13–28). Barcelona, Spain: Departament de Didàctica de la Llengua de la Literatura i de les Ciències Socials, Universitat Autònoma de Barcelona.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (Eds.) (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain.* New York: David McKay.

Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences.* London: Lawrence Erlbaum Associates.

Bowlby, J. (1969). *Attachment and loss: Vol. 1. Attachment.* New York: Basic Books.

Downey, C. J., Steffy, B. E., English, F. W., Frase, L. E., & Poston Jr., W. K. (Eds.). (2004). *The three-minute classroom walk-through: Changing school supervisory practice one teacher at a time*. Thousand Oaks, CA: Corwin Press.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Abe, H., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software, 39*(12), 1–25. doi: 10.18637/jss.v039.i12

Doherty, K. M., & Jacobs, S. (2013). State of the States 2013: Connect the Dots--Using Evaluations of Teacher Effectiveness to Inform Policy and Practice. *National Council on Teacher Quality*.

Downer, J. T., Stuhlman, M., Schweig, J., Martínez, J. F., & Ruzek, E. (2015). Measuring effective teacher-student interactions from a student perspective: A multi-level analysis. *Journal of Early Adolescence*, *35*(5-6), 722-758.

Dutch Inspectorate of Education (2016). *De staat van het onderwijs 2014-2015* [*The state of education in the Netherlands 2014-2015*]. De Meern, Inspectie van het Onderwijs.

Ferguson, R. F., & Danielson, C. (2014). How framework for teaching and tripod 7Cs evidence distinguish key components of effective teaching. In Thomas J. Kane & Kerri A. Kerr (Eds.), *Designing teacher evaluation systems (*pp. 98-143). San Francisco, CA: Wiley and Sons, Inc.

Fuller, F. (1969). Concerns of teachers: A developmental conceptualization. *American Educational Research Journal, 6*, 207–226.

Gelman A., Su Y.-S., Yajima M., Hill J., Pittau M. G., Kerman J., et al. (2015). *ARM: Data analysis using regression and multilevel/ hierarchical models*. R package version 1.8-6 2015: URL: https://cran.r-project.org/web/packages/arm/arm.pdf

Goe, L., & Croft, A. (2009). *Methods of evaluating teacher effectiveness*. Washington, DC: National Comprehensive Center for Teacher Quality.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher, 41*(2), 56-64.

Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A. J., Jones, S. M., Brown, J. L., Cappella, E., Atkins, M., Rivers, S. E., Brackett, M. A., & Hamagami, A. (2013). Teaching through interactions: Testing a developmental framework of teaching effectiveness in over 4,000 classrooms. *Elementary School Journal, 113,* 461–487.

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.

Huberman, M. (1993). *The lives of teachers.* New York: Teachers College Press.

Isoré, M. (2009). *Teacher Evaluation: Current Practices in OECD Countries and a Literature Review.* OECD Education Working Papers, No. 23. OECD Publishing (NJ1).

Kane, T. J., & Staiger, D. O. (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.

Kolen, M. J., & Brennan, R. L. (2014). *Statistics for social and behavioral sciences. Test equating, scaling, and linking: Methods and practices*. New York: Springer Science+ Business Media.

Korpershoek, H., Harms, T., de Boer, H., van Kuijk, M., & Doolaard, S. (2016). A meta-analysis of the effects of classroom management strategies and classroom management programs on students' academic, behavioral, emotional, and motivational outcomes. *Review of Educational Research, 86*(3), 643–680.

Kyriakides, L., Creemers, B. P., & Panayiotou, A. (2018). Using educational effectiveness research to promote quality of teaching: The contribution of the dynamic model. *ZDM, 50*(3), 381-393.

Mair, P., & Hatzinger, R. (2007). Extended Rasch modelling: The eRm package for the application of IRT models in R. *Journal of Statistical Software 20*(9), 1–20.

Marsh, H. D. (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). Dordrecht, The Netherlands: 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,* 1–23, doi:10.1007/s10984-016-9215-8

Muijs, D., & Reynolds, D. (2003). Student background and teacher effects on achievement and attainment in mathematics: A longitudinal study. *Educational Research and Evaluation, 9*(3), 289–314. doi: 10.1076/edre.9.3.289.15571

Murray, H. G. (1983). Low-inference classroom teaching and student ratings of college teaching effectiveness. *Journal of Educational Psychology, 75*(1), 138–149.

Polikoff, M. S. (2015). The stability of observational and student survey measures of teaching. *American Journal of Education, 121*(2), 183-212. doi: 10.1086/679390

Ponocny, I. (2001). Non-parametric goodness-of-fit tests for the Rasch model. *Psychometrika 66*(3), 437–460.

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.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Nielsen & Lydiche.

Reis, S. M., McCoach, D. B., Little, C. A., Muller, L. M., & Kaniskan, R. B. (2011). The effects of differentiated instruction and enrichment pedagogy on reading achievement in five elementary schools. *American Educational Research Journal*, *48*(2), 462–501.

Rosenshine, B. (1995). Advances in research on instruction. *Journal of Educational Research, 88*(5), 262–268.

Ryan, R., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*, 68–78.

Scriven, M. (1981). Summative teacher evaluation. In J. Millman (Ed.), *Handbook of Teacher Evaluation*. Beverly Hills, CA: Sage Publications.

Steele, J. L., Hamilton, L. S., & Stecher, B. M. (2010). Incorporating Student Performance Measures into Teacher Evaluation Systems. Technical Report. *Rand Corporation*.

Tas, T., Houtveen, T., Van de Grift, W., & Willemsen, M. (2018). Learning to teach: Effects of classroom observation, assignment of appropriate lesson preparation templates and stage focused feedback. *Studies in Educational Evaluation*, *58*, 8-16.

Van de Grift, W. J. C. M., Van der Wal, M., & Torenbeek, M. (2011). Ontwikkeling in de pedagogisch didactische vaardigheid van leraren in het basisonderwijs [The development of teaching skills]. *Pedagogische Studiën, 88*(6), 416-432.

Van de Grift, W., 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.

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., 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. (2017). Individual differences in teacher development: An exploration of the applicability of a stage model to assess individual teachers. *Learning and Individual Differences*, *58*, 46-55.

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.

Wentzel, K. R. (2002). Are effective teachers like good parents? Teaching styles and student adjustment in early adolescence. *Child Development, 73*(1), 287–301.

Table 1. *Six domains and corresponding items from the MTQ and ICALT*

Instrument	**Domain**	**Example item**
MTQ	Safe learning climate	My teacher ensures that I feel relaxed in class.
ICALT	Safe learning climate	This teacher creates a relaxed atmosphere.
MTQ	Efficient classroom management	My teacher applies clear rules.
ICALT	Efficient classroom management	This teacher ensures effective class management.
MTQ	Clear and structured explanation	My teacher uses clear examples.
ICALT	Clear and structured explanation	This teacher explains the subject matter clearly.
MTQ	Active teaching methods	My teacher encourages me to think.
ICALT	Active teaching methods	The teacher asks questions that encourage students to think.
MTQ	Teaching learning strategies	My teacher explains how I should study something.
ICALT	Teaching learning strategies	This teacher asks students to reflect on approach strategies.
MTQ	Differentiating in instruction	My teacher knows what I find difficult.
ICALT	Differentiating in instruction	This teacher adapts processing of subject matter to student differences

Table 2. C*umulative item ordering (least to most difficult teaching practices) (O = observation item; S = survey item)*

Domain	Item	Description: This/My teacher…	b	SE
climate	O1	shows respect for students in behavior and language	-2.25	.40
climate	S21	treats me with respect	-1.47	.14
climate	O2	creates a relaxed atmosphere	-1.38	.31
management	S20	prepares his/her lesson well	-1.18	.14
management	O7	ensures effective class management	-1.09	.28
climate	O3	supports student self-confidence	-1.05	.28
climate	S40	helps me if I do not understand	-.94	.13
instruction	O9	explains the subject matter clearly	-.92	.28
climate	S6	answers my questions	-.89	.13
management	O5	ensures that the lesson runs smoothly	-.79	.26
climate	O4	ensures mutual respect	-.69	.26
instruction	O14	gives well-structured lessons	-.64	.26
management	S3	makes clear what I need to study for a test	-.61	.13
management	O8	uses learning time efficiently	-.56	.25
management	S19	makes clear when I should have finished an assignment	-.52	.13
climate	S8	ensures that I treat others with respect	-.46	.13
climate	S1	ensures that others treat me with respect	-.44	.13
instruction	S13	explains the purpose of the lesson	-.35	.13
instruction	S24	uses clear examples	-.33	.13
management	S23	ensures that I pay attention	-.26	.13
management	S26	applies clear rules	-.15	.12
management	O6	checks during processing whether students are carrying out tasks properly	-.10	.23
management	S2	ensures that I use my time effectively	-.10	.12
instruction	O15	clearly explains teaching tools and tasks	-.08	.23
instruction	O10	gives feedback to students	-.03	.23
instruction	O11	involves all students in the lesson	.03	.22
instruction	S39	Involves me in the lesson	.06	.12
instruction	O13	encourages students to do their best	.11	.22
instruction	S33	ensures that I know the lesson goals	.12	.12
activation	S17	encourages me to think for myself	.39	.12
activation	O19	asks questions that encourage students to think	.50	.21
activation	S12	ensures that I keep working	.53	.12
activation	O16	uses teaching methods that activate students	.58	.21
activation	S30	stimulates my thinking	.68	.12
activation	O21	provides interactive instruction	.71	.21
instruction	O12	checks during instruction whether students have understood the subject matter	.74	.21
activation	O20	has students think out loud	.84	.20
differentiation	S25	connects to what I am capable of	.89	.12
differentiation	S34	checks whether I understood the subject matter	1.15	.12
learning strategies	O30	encourages students to apply what they have learned	1.28	.20
learning strategies	S16	teaches me to check my own solutions	1.52	.12
learning strategies	O31	encourages students to think critically	1.64	.20
differentiation	S36	knows what I find difficult	1.68	.12
differentiation	O23	checks whether the lesson objectives have been achieved	1.96	.20
learning strategies	O28	encourages the use of checking activities	2.16	.20
learning strategies	O29	teaches students to check solutions	2.21	.20
differentiation	O25	adapts processing of subject matter to student differences	2.60	.20
differentiation	O26	adapts instruction to relevant student differences	2.77	.20

Figure 1. Staged progression of teacher development of effective teaching

Notes: Checks indicate that the teaching behaviors associated with this stage are observed, crosses indicate the behaviors are not observed.

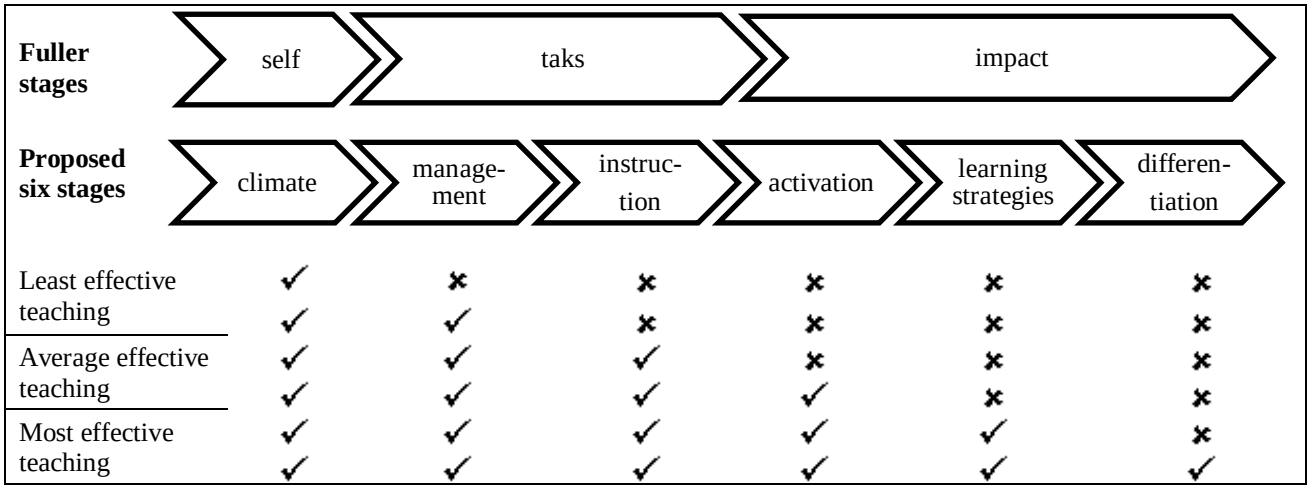

Figure 2. Students and observer disagree about the teacher's A teaching quality, yet all item responses fit the predicted stage pattern. Check-boxes indicate the teacher is rated positively (=1), crosses indicate negative ratings (=0).

		Climate	Manage-ment	Explan-ation	Acti-vating	Learning strategies	Differen-tiation	total score
Student 1	Teacher A	✓	✗	✗	✗	✗	✗	1
Student 2	Teacher A	✓	✓	✗	✗	✗	✗	2
Student 3	Teacher A	✓	✓	✓	✗	✗	✗	3
Student 4	Teacher A	✓	✓	✓	✓	✗	✗	4
Student 5	Teacher A	✓	✓	✓	✓	✓	✗	5
Observer 1	**Teacher A**	✓	✓	✓	✓	✓	✗	5

Figure 3

Illustration of cumulative item order of survey items and classroom observation items and an example of misfit (teacher G).

	Climate		**Management**		**Explanation**		**Activation**		**Learning strategies**		**Differentiation**	
	Observation item The teacher ensures a relaxed atmosphere	Survey item My teacher ensures that I feel relaxed in class	Observation item The teacher uses learning time efficiently	Survey item My teacher makes sure that I use my time effectively	Observation item The teacher involves all students in the lesson	Survey item My teacher involves me in the lesson	Observation item This teacher asks questions that encourage students to think	Survey item My teacher motivates me to think	Observation item The teahcer asks students to reflect on approach strategies	Survey item My teacher asks me how I am going to learn the content of the lesson	Observation item The teacher adapts processing of subject matter to student differences	Survey item My teacher knows what I find difficult
Teacher A	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Teacher B	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗
Teacher C	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗
Teacher D	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗
Teacher E	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗
Teacher F	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Teacher G	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗

Figure 4. GoF plot for subgroups with the poorest fit (left) and best fit (right).