# Chapter 3
# Teacher evaluation on the basis of student ratings

# Abstract

This study reports on the development of a teacher evaluation instrument, based on students' observations of teaching practices, that exhibits cumulative ordering in terms of the complexity of teaching practices. The study integrates theory on teacher development with theory on teacher effectiveness and applies a cross-validation procedure to verify whether effective teaching practices have a cumulative order. The resulting teacher evaluation instrument comprises 32 effective teaching practices with cumulative ordering in terms of complexity. This ordering aligns with prior teacher development research. It also represents a valuable extension, in that the instrument can provide feedback about a teacher’s current phase of development and advice for improvement

Modified version of the article: Van der Lans, R. M., Van de Grift, W., J., C., M., & Van Veen, K. (2016).Developing a teacher evaluation instrument to provide formative feedback using student ratings of teaching acts. *Educational Measurement: Issues and Practice, 34*(3), 18-27.

## 3.1 Introduction

Many Western countries seek to improve education by adopting revised teacher evaluation policies. The drivers of this shift are value-added teacher evaluations (e.g., Firestone, 2014), which are designed to describe the extent to which a teacher has contributed to student achievement gains in a school year. However, value-added evaluations can only inform teachers about their gains; they shed light on neither why students obtained that gain nor how they could improve their gains (Darling-Hammond, Amrein-Beardsley, Heartel, & Rothstein, 2012; Firestone, 2014; Hill, Kapitula, & Umland, 2011). Therefore, current consensus holds that other evaluation instruments are required to complement value-added teacher evaluations.

This consensus has shifted research attention toward further development of classroom observation instruments and student survey instruments as means for evaluation (e.g., Danielson, 2013; Hill et al., 2011; Bill & Melinda Gates Foundation, 2012). Although these instruments are effective in providing more precise information about what a teacher does inside the classroom, such information does not automatically translate into formative feedback (i.e., information about how to develop and improve further). The provision of feedback would require connecting teacher evaluation instruments with teacher development theories. Available teacher development research indicates that the process of becoming an expert teacher follows specific, sequentially or cumulatively ordered phases (Berliner, 2001; Fuller, 1969). Despite widespread acceptance of these theories, the field lacks an evaluation instrument that can provide feedback about which phase of development a teacher has reached and which teaching skills should be considered next for ongoing teacher training, reflection, and self-study. We propose a teacher evaluation instrument that exhibits cumulative ordering and that can provide formative feedback to teachers about their current phase in development.

## 3.2 Theoretical background

The theoretical background is structured in three parts: We consider and summarize teacher development theories; then relate them to key findings about teacher effectiveness; finally, we consider the pros and cons of two evaluation methods; student ratings and classroom observations.

### 3.2.1 Theories of teacher development

Theories of teacher development describe progressive changes in teacher concerns (Fuller, 1969) as well as a progressive development from novice to expert (Berliner, 2001). From such works, we seek to define an ordering that parallels and can be integrated with findings from teacher effectiveness literature. However, we acknowledge though that theories of teacher development traditionally focus on (sequential stages in) teacher *cognition*, rather than teacher *behavior*, which is the focus in teacher effectiveness research. Therefore, our exploration relies on the presumption that teachers' (cognitive) concerns partially reflect observable difficulties and changes they encounter in their teaching. In addition, we note that theory on teacher development, and in particular Fuller's theory, have stimulated two different strands of research (Conway & Clark, 2003); one dedicated to the description of the developmental dynamics of teaching, and one dedicated to the evaluation of teacher concerns in the context of innovation and reform. This paper contributes and is connected with the former; description and measurement of the development of teaching.

Fuller's (1969) theory of teacher concerns is among the first to describe teacher development. It features a relatively simple, three-stage model in which teachers first are concerned with the self, before they turn their attention to tasks, and finally toward students and the impact of their teaching (Conway & Clark, 2003). Concerns for the self center on issues of authority, respect, status, and relationships. Concerns with tasks involve classroom management and content adequacy. Concerns with the impact of student learning pertain to teachers' ability to specify objectives for students, understand students' capacities, and identify their own contributions to students' difficulties (Fuller, 1969).

Teacher development in Fuller's first two stages, in particular, is well documented. Berliner (2001) describes teachers' growth from novice to expert. For novice teachers, Berliner highlights the importance of developing classroom routines for management and instruction (i.e., tasks). The life-cycle teacher career model (Steffy & Wolfe, 2001) describes six phases, ranging from novice to emeritus, and predicts that teachers who have successfully completed their teacher education begin by developing routines for lesson preparation and achieving reciprocal respect (i.e., task and self). Schafer, Stringfield, and Wolfe (1992) conclude, on the basis of a two-year longitudinal study, that classroom management and basic instruction are among the first teaching skills acquired by teachers (i.e., task).

Regarding the third stage, the impact of student learning, current understanding about its development is limited. The few works exploring the development of more experienced teachers conclude that, in contrast with the relatively homogeneous development of more elementary stages, acquiring skill in the more complex stages is much more varied among teachers, and some teachers never acquire them (e.g., Berliner, 2001; Huberman, 1993).

The discussion has also focused on the rigidity of the proposed stages. Fuller's theory has been characterized as "Perhaps the most classic of stage theories in that it was meant to be relatively invariant, sequential and hierarchical" (Richardson & Placier, 2001, p. 910). In contrast, Berliner (2001), Steffy and Wolfe (2001), and Huberman (1993) suggest a more tentative heuristic interpretation in terms of phases in teacher development. In their view, teachers can develop competence at any time during any phase, and yet at any moment also be grouped into one best-fitting phase. This tentative heuristic approach has the advantage of being less restrictive when describing individual differences in the development of teaching skill, but at a cost: Because it does not exclude any developmental trajectory, information about current teaching does not reveal the most logical steps for further development and improvement. As Richardson and Placier (2001, p. 913) conclude, "the use of a very flexible approach to stages or phases may have taken us so far from the original concept of a stage theory that the usefulness of the work must be rethought." In contrast, Fuller's invariant, hierarchical approach restricts the individual variation in development of teaching skill, but—if valid—it has the potential to inform an individual teacher about logical steps for ongoing training, reflection, and self-study.

### 3.2.2 Teacher effectiveness literature and development in teaching skill

Several reviews and meta-analyses address the relation between teaching practices and student achievement (Hattie, 2009; Kyriakides, 2013; Marzano, 2003), and though they use different labels, they show consistently that similar categories of teaching practices enhance student achievement. We consider six broad domains of teaching practices that can be observed within classrooms: creating a safe learning climate, efficient classroom management, quality of instruction, student activation, teaching learning strategies, and differentiation (the questionnaire items are included in Appendix C [English translation] and E [Dutch version]). Van de Grift (2014) provides an extensive literature review to account for the six domains. In addition, Maulana, Helms-Lorenz, & Van de Grift (2015)

describe connections between these six domains and the classroom assessment scoring system (CLASS) and the framework for teaching (FFT) observation protocol, both of which are currently employed in the Measures of Effective Teaching (MET) project. They conclude that the six domains coincide with all the clusters of the FFT and CLASS.

Table 3.1 compares the six domains with the seven Cs of the Tripod survey (Bill & Melinda Gates Foundation, 2012), a student questionnaire employed in the MET project.

**Table 3.1.**

Framework formulating the assumed relations between the six domains of teaching acts and Fuller's three-stage model. Also, the comparison of the Tripod survey and the six domains.

| **Fuller Stage** | **Domain** | **Tripod Survey Factors** |
|---|---|---|
| *Self:* concerns for their authority, respect, status, and relationships | *Safe Learning Climate:* relation between teacher and class. | *Caring:* Encouragement and support<br>*Confer:* Students sense their ideas are respected. |
| *Task:* concerns about how to mobilize resources. | *Efficient Classroom Management:* overall order in the classroom.<br>Quality of Instruction: basic explanation of lesson topics, the overall lesson structure, and connections among lesson parts. | *Control:* Culture of cooperation and peer support<br>*Clarifying:* Teaching should evoke a sense that success is feasible<br>*Consolidate:* Ideas get connected and integrated |
| *Impact:* concerns for their ability to specify objectives, and how to partial out own contributions to students' difficulties. | *Student Activation:* motivating students to think about the topic.<br>*Teaching Learning Strategies:* efforts to teach students how to learn.<br>*Differentiation:* demonstrations of sensitivity and flexibility to meet individual students' learning problems and needs. | *Challenge:* Press for effort, perseverance, and rigor<br>*Captivating:* Learning seems interesting and relevant |

The Tripod survey is clustered into seven factors—caring, controlling, clarifying, challenging, captivating, conferring, and consolidating—that measure how students experience the teacher's behavior. As Table 3.1 shows, the overall impression is that the seven Cs coincide with four domains: safe learning climate, efficient classroom management, quality of instruction, and activating students. The learning strategies and differentiation domains appear relatively unique to our framework.

In addition, Table 3.1 notes possible connections between the six domains and Fuller's three stages of teachers' concerns. We acknowledge that these connections are to some extent speculative, but they may contribute to an understanding of the six domains in terms of progressive stages. Our speculations are based on some recent empirical studies (Kyriakides, Creemers, & Antaniou, 2008; Van de Grift, Van der Wal & Torenbeek, 2011) that indicate a cumulative ordering of teacher practices, from less to more complex, which may reflect teaching development. Kyriakides, Creemers, and Antaniou (2008) group teaching practices into five types and find a cumulative ordering that gradually moves from actions associated with direct teaching to more advanced actions involving new teaching approaches and differentiation. Van de Grift, Van der Wal & Torenbeek (2011) analyze classroom observations performed by trained colleagues in elementary education of the identical six domains of effective teaching practices. This study found they are cumulatively ordered, from a safe learning climate to efficient classroom management to quality of instruction to student activation and finally to differentiation and then learning strategies.

### 3.2.3 Evaluation method: student ratings

The success of an evaluation instrument depends on its ability to present feedback to individual teachers about their teaching. This criterion creates some different and unusual demands. Unlike the conventional goal of empirical research—to generalize across people—our focus is on generalizing across situations in which a person acts. Furthermore, the chosen method ideally has low implementation costs but still provides feedback that is informative about a relatively wide range of situations. With these considerations, we discuss the advantages and disadvantages of two observational methods: classroom observations and student ratings.

**Classroom observations.** A classroom observer may be a trained assessor or someone with extensive experience observing classrooms. The principal advantage of

classroom observation is that the observer is not involved in any way in the lessons. Ideally, well-trained observers evaluate teachers using a similar norm and therefore should be more objective (Muijs, 2006). However, a single observation cannot reflect the teacher's average performance over a larger set of situations. To achieve reliable estimations of performance across time, some studies recommend three to six classroom observations (e.g., De Jong & Westerhof, 2001; Hill, Charalambous, & Kraft, 2012; Van der Lans et al., 2016). Another disadvantage of this method is the potential for observer bias. If only one observer evaluates the teacher on multiple occasions, those observations could reflect the observer's prejudices and personal values; interaction effects between observers and teachers also could clutter the evaluation results. The solution would be to have multiple observers assess the teacher (Peterson, 2000). Overall then, classroom observation offers the advantages of an objective, outside perspective, but it requires the use of multiple trained observers who observe each teacher on three to six occasions in each class. For schools to adopt classroom observations for their teacher evaluations, the costs would likely be enormous, while the benefits yet remain uncertain.

**Student ratings.** Researchers and teachers have long been suspicious of student ratings. Because students are closely involved in the lessons, they are not independent or objective raters. However, most recent research indicates that student ratings can provide trustworthy, valid insights for teacher evaluation (Marsh, 2007). An advantage of student ratings is that they usually span many observers at once, thereby substantially decreasing observer bias (Marsh, 2007; Richardson, 2005). In addition, research shows that students ratings vary primarily as a function of the teacher's teaching skill (Benton & Cashin, 2012; Richardson, 2005). Furthermore, student ratings tend to be stable over time (Benton & Cashin, 2012), which suggests that students rate teaching practices according to their average perception across all previous encounters. These advantages make student ratings considerably more cost effective than classroom observations. Concerns with student ratings mostly involve the potential for bias. Researchers have directed considerable attention to bias due to students' expectations about their grades (i.e., whether students favor lenient graders) and due to students' prior interest in the subject matter (i.e., whether students misattribute their own subject matter interest to be caused by the teacher), but these biases are generally small (Benton & Cashin, 2012; Marsh, 2007; Richardson, 2005). More profound concerns relate to student expertise; younger students in particular may not be aware of valuable information required to evaluate teachers (Peterson, 2000).

Furthermore, students are not trained observers, and compared with classroom observers, they have relatively little experience with differences in teaching. In summary, student ratings offer a relatively cost-effective evaluation method, because they are unpaid evaluators and require few evaluation moments, but the evaluations reflect what students expect from the teacher, not a trained preset, standardized norm.

Against this background, we address the following research questions: Can student questionnaire ratings of effective teaching practices be ordered cumulatively? And; How may the development of such a scale contribute to the knowledge about teacher development?

## 3.3 Method

### 3.3.1 Sample

The sample for this study consisted of 2,262 student ratings, obtained from a school for secondary education in the Netherlands (student ages: 12–18 years). Female students constituted 53.1% of the student sample (1,200). The school offers vocational, higher vocational, and pre-university education. Students judged 68 teachers working at the school. The study included teachers from all subjects except Physical Education. Teaching experience ranged from 0 to 43 years, with an average of 16 years.

### 3.3.2 Measurement instrument

The applied version of the "My Teacher" questionnaire includes items reflecting 59 effective and observable teaching practices, such as "This teacher knows what I'm able to do" or "This teacher ascertains that I understand the subject matter taught" (see also Appendix C). The same questionnaire has been applied in research on induction, using a sample of beginning secondary education teachers (< 3 years teaching experience) (Maulana et al., 2015). The original student questionnaire had four response categories; 1 = "weak", 2 = "more weak than strong", 3 = "more strong than weak", and 4 = "strong". These four original response categories were dichotomized where 1 and 2 were considered "weak" and 3 and 4 were recoded as "strong". We deliberately chose for a dichotomous response coding, because in the Rasch model its interpretation is more straightforward and though the feedback is more easily explained to teachers. Simplicity is perceived key to implementation.

Nevertheless, we checked whether the dichotomization did not lead to an unacceptable loss of information. For this purpose a Graded Response Model (GRM) was applied. The GRM is identical to the Rasch model, except that it can handle multiple response categories. Using the GRM, the latent variable teaching scale had a range of 8.95, compared to a range of 7.20 if using the binary Rasch Model. The models had identical averages GRM: $M$ = 1.98 and Rassh model $M$ = 2.01 and their Spearman rank correlation (rho) is .84. Together these results give the impression that the dichotomization did not lead to an unacceptable loss of information.

### 3.3.3 Design and missing values

In the nested design, the aggregate level identifies 84 unique teacher–class combinations. This number exceeds the number of teachers in the data set because for some teachers, ratings were available from two classes, resulting in two unique combinations. Of the 131,458 item responses given, 2,016 were reported missing, a 1.5% rate of missing values. We considered these missing values to be missing at random (MAR).

### 3.3.4 Cross-validation procedure

The method relied on a cross-validation procedure for which the complete sample was split into development and validation samples. The complete sample counted 2,262 student ratings. We established a development sample by randomly selecting 10 students from each teacher–class combination ($n_{development}$ = 840). This development sample served to calibrate the measurement instrument.

To establish the validation sample, we randomly selected another 10 students from each teacher–class combination ($n_{validation}$ = 750). The validation sample was slightly smaller than the development sample because a few classes contained fewer than 20 students, so fewer than 10 students remained for the validation sample; six teacher–class combinations had fewer than 6 student ratings left to include in the validation sample. To limit sample imbalance, we excluded these combinations, such that the validation sample also featured six fewer teachers than the development sample.

In total, 1,590 students are included in the development and validation samples. The other 672 students were omitted. Subsamples did not differ in student total test scores ($F$(2, 2214.58) = .27, $p$ = .79) or in student age ($F$(1, 2193.89) = .20, $p$ = .66), though they did

differ slightly on student gender ($\chi^2(2, n = 2{,}226) = 6.90$, $p = .03$). The omitted sample had 57.8% girls, while the two randomly selected samples; 51.3% and 52.4%.

### 3.3.5 Model specification



Our research question pertains to whether we can find a cumulative order for effective teaching practices. To address it, we apply the Rasch model, generally considered the most appropriate model to test for cumulative item ordering (Bond & Fox, 2007). The Rasch model relies on three assumptions (DeMars, 2010):

4. *Parallel item characteristic curves (ICCs).* This assumption states that each teaching practice can discriminate equally among levels of teaching skill.
5. *Unidimensionality.* This assumption states that student responses can be ascribed to a single latent construct: teaching skill.
6. *Local independence.* This assumption states that the residuals of item pairs are uncorrelated.

We deliberately chose the strict one-parameter item response theory (IRT) model (i.e., the Rasch model) instead of the two-parameter IRT model. We view the two-parameter IRT model as an effective option to develop latent measurement scales, but it cannot be applied to test for cumulative ordering, as is examined in this study (Bond & Fox, 2007).

The Rasch model can be understood as a generalized linear mixed model specifying two components (De Boeck et al., 2011):

$$\log\left(\frac{P_{\mathrm{ij}}}{1-P_{\mathrm{ij}}}\right) = \sum_{\mathrm{j=1}}^{\mathrm{J}} \theta_{\mathrm{p}} Z_{(\mathrm{p,i})\mathrm{j}} + \sum_{\mathrm{k=1}}^{\mathrm{K}} b_{\mathrm{i}}\, X_{(\mathrm{p,i})\mathrm{k}}$$

We refer to the first of these components as the “structural model” and the second as the “measurement model.” Interactions between the components suggest model violations. The plus sign signals that $b_i$ should be interpreted as item easiness. If—as in our case—the design is nested and a third component is added to this equation, we must specify whether the third component is nested within the structural model or within the measurement model. In this study, we view students as nested in teachers, and they together define the structural model. Because items are not nested, their fit is assessed by application of the regular

single-level item fit statistics. As we discuss at the end, we view this approach as defendable yet not entirely satisfactory.

### 3.3.6 Data analysis

The analyses consist of two sections: (1) validation of the measurement model and (2) examination of the structural model. The validation of the measurement model is further subdivided in two subsections: development and validation phases.

**Validation of the measurement model.** In the development phase, we tested for item fit with the three Rasch model assumptions. We excluded from further analysis any item that did not meet any one of these assumptions. Test included are; Andersen (1973) likelihood ratio (LR) test to evaluate the assumption of parallel ICC, exploratory factor analysis (EFA) to evaluate the assumption of unidimensionality, and Ponocny's (2001) nonparametric $T_1$ and $T_{1m}$ to evaluate the assumption of local independence.

In the validation phase, we reassessed the fit of the remaining items to ensure that the teaching practices described by the items had not been selected on the basis of chance. This second phase is directed at validation, not item selection. The validation involved identical tests with exception of the EFA; we consider confirmative factor analysis (CFA) more appropriate for validation.

**Structural model: An exploration of measurement reliability.** In this section, the results involve the measurement reliability and marginal standard error of measurement (SEM). Following Raju, Price, Oshima, and Nering (2006), we estimate the group reliability for teachers ($\rho_{(\theta\theta')T}$) and students ($\rho_{(\theta\theta')S}$). Analogous to Patz, Jucker, Johnson, and Mariano (2002), we turn to the hierarchical structure and explore how raw scores are translated into different values of θ scale and its associated SEM. However, unlike Patz et al. (2002) but consistent with Brennan (2004), we do not interpret the rater facet as constituting bias or rater severity. Variation in students' ratings is equally interesting and may ultimately prove useful in informing teachers about possible steps to improve their teaching with regard to particular target students; however, the scope of this discussion transcends the primary goal of this article: to develop a Rasch-scaled student rating instrument for teacher evaluation.

### 3.3.7 Software

The data analysis procedure relied on R and Mplus version 7 (Muthen & Muthen, 1998–2012). In R we installed the eRm R-package (Mair & Hatzinger, 2007), which uses a conditional maximum likelihood algorithm to estimate the item fit statistics. Mplus applies a robust weighted least squares estimator algorithm to estimate item fit. The nested components of the structural model were estimated with the R package lme4 version: 1.1-7 (Bates Maechler, Bolker, & Walker, 2014).

## 3.4 Results

We begin this section by presenting the results for measurement model. Starting with the instrument calibration in the development sample, then a reexamination of item fit in the validation sample and ending with a presentation of our proposed evaluation instrument. The result section then turns to the structural model and explores measurement reliability.

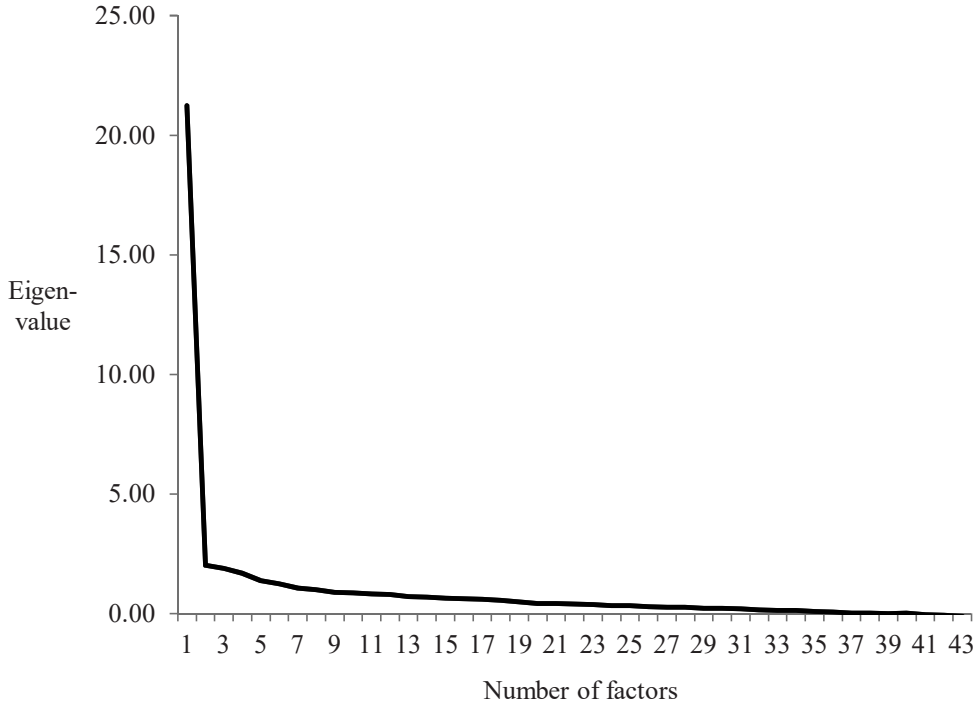
### 3.4.1 Development sample

**Parallel ICC.** Anderson (1973) proposes an LR test of parallel ICCs, splitting observed data into two subgroups: one that scores low on the measured latent trait (i.e., low teaching skill) and another that scores high on it (i.e., high teaching skill). The LR test then compares the deviance in the log-likelihood ratios of both groups against a chi-square distribution. We performed the median as the split criterion. The LR test revealed that not all 59 items achieved parallel ICC ($\chi^2 = 286.10$, $df = 58$, $p = .00$). Therefore, we excluded the teaching practice that resulted in the greatest decrease in the chi-square value over repeated rounds, until 43 of the initial 59 items remained; on average, they exhibited parallel ICC ($\chi^2 = 54.50$, $df = 42$, $p = .09$).

**One-dimensionality.** The one-dimensionality assumption is difficult to (dis)confirm (DeMars, 2010). All measurement instruments are, to some extent, multidimensional, and we can only test whether one-dimensionality is defensible. A common strategy uses factor analysis, which suggests that, provided one-dimensionality holds, the best factor solution of the correlations among the 43 items should be a one-factor solution. We used tetrachoric correlations, because Pearson phi correlation coefficients can prompt high loadings for ratings with similar difficulty (DeMars, 2010). The eigenvalues of the EFA, as plotted in Figure 3.1, suggest a one-factor solution. The first eigenvalue (21.23) is considerably larger than the second (2.01) and third (1.89) eigenvalues.

**Figure 3.1**

Scree plot of the exploratory factor analysis using the tetrachoric correlations.




**Note.** The y-axis shows the eigenvalue, and the x-axis indicates the number of factors.

**Local independence.** To test the local independence assumption, we used Ponocny's (2001) $T_1$ and $T_{1m}$. Rasch (1960) was especially concerned about this third assumption of his model and originally proposed, but never completed, a nonparametric test to assess model fit. Ponocny's (2001) family of T-statistics implements some of Rasch's original design. The T-statistics specify each an one-tailed directional alternative hypothesis, which increases their power considerably. The $T_1$ statistic evaluates violations of local independence due to increasing (i.e., positive) residual correlations, and the $T_{1m}$ statistic evaluates violations due to decreasing (i.e., negative) residual correlations.

Chance should have an important position in evaluating the *T*-statistics results (I. Ponocny, personal correspondence, September 30, 2014). The *T*-statistics pair every item with 42 other items. Therefore, a criterion of two violations per item would reflect an alpha criterion of .05. However, their considerable power together with the slight overlap in item content (both within and between domains) and students' differential grammar ability,

makes that some additional violations are almost inescapable and may be tolerated. On the basis of these considerations we decided to set a more lenient criterion of 5 violations.

Items 5 ("my teacher explains well") and 51 ("this teacher makes sure I understand his/her explanation") together yielded 33 of the total 109 violations for $T_1$. Moreover, item 15 ("my teacher asks questions that make me think") alone accounts for 25 violations for $T_{1m}$. Continuing with the calibration, we deleted additional items over repeated rounds, starting with the item that accounted for the most violations. After excluding 13 items, the 32 remaining items combined for 37 violations due to increasing correlations and 28 violations due to decreasing correlations and no item accounted for more than 5 violations.

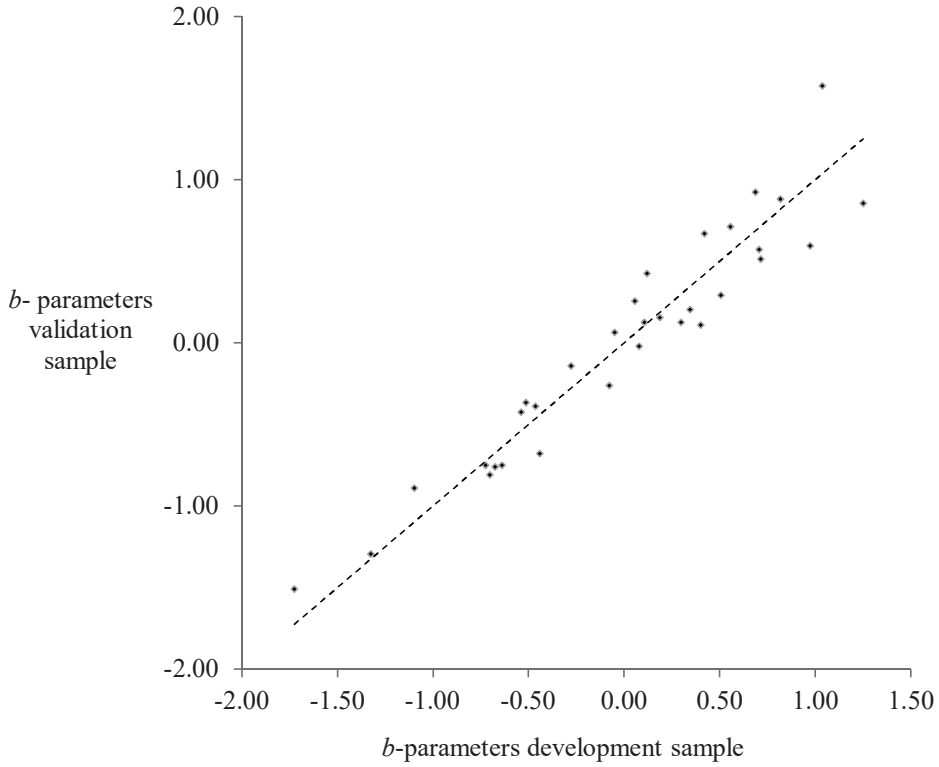
### 3.4.2 Validation sample

We reexamined the fit of the 32 items with each of the Rasch model assumptions using the validation sample ($n_{\text{validation}} = 750$). The Andersen LR test confirmed that, on average, all items achieved parallel ICC ($\chi^2 = 36.90$, $df = 31$, $p = .22$). A CFA, applied to reexamine the one-dimensionality assumption, showed that the one-factor model fit the data well (root mean square error of approximation = .029, confirmatory fit index = .96, Tucker–Lewis index = .96). The scree plot confirmed that the one-factor solution was defensible. Finally, with regard to local independence, Ponocny's (2001) $T_1$ indicated that four items had more than 5 violations, and $T_{1m}$ indicated that three items had more than 5 violations (see Table 3.1). In total, 7 of the 32 items failed to meet the local independence criterion in the validation analysis, but these 7 violations did not seem to cluster around any particular domain. Overall, we consider these results encouraging.

In addition, we examined the invariance of the item ($b$) parameters between the development and validation samples. Figure 3.2 shows the goodness-of-fit plot. The 32 dots indicate the 32 items, the dashed line reflects the perfect invariance between samples (i.e., the zero-difference score). The deviations from the dashed line indicate deviations from item invariance. The goodness-of-fit plot shows that—with the exception of item 12 ("my teacher treats me with respect")—the item parameters can be considered invariant between samples.

**Figure 3.2**

Goodness-of-fit plot visualizing item parameter invariance between the development and validation samples.




**Note.** The x-axis gives the item complexity (*b*-) parameters for the development sample. The y-axis gives the item complexity parameters for the validation sample. The dashed line represents complete invariance, and deviations from the dashed line indicate deviations from sample invariance.

### 3.4.3 Final questionnaire

We present the established scale in Table 3.2. The *b* coefficients indicate the difficulty (i.e., here complexity) of the teaching practice, such that low values signify teaching practices with less complexity. Because these 32 teaching practices fit our criteria for cumulative ordering, it follows that the practices with higher *b* coefficients could have been rated "often" by students only if (most) teaching practices with lower *b* parameters also were rated "often". Thus, the less complex teaching practices can be considered prerequisites for more complex teaching practices.

**Table 3.2**

Fuller stage, domain, and complexity (b) of 32 teaching practices (n = 1,590)

| stage | domain | teaching practice | *b* | $SE_{(b)}$ |
|---|---|---|---|---|
| self | climate | treats me with respect. | –1.32 | .176 |
| task | management | prepares his/her lesson well. [b] | –.98 | .166 |
| self | climate | ensures that others treat me with respect. | –.72 | .160 |
| self | climate | answers my questions. | –.69 | .159 |
| self | climate | ensures that I treat others with respect. | –.67 | .159 |
| task | management | makes clear what I need to study for a test. | –.65 | .158 |
| task | management | helps me if I do not understand or am unable to do something.[a] | –.56 | .156 |
| task | instruction | uses clear examples.[a] | –.55 | .156 |
| task | management | ensures that I know what to do. | –.49 | .155 |
| task | management | ensures that I behave well. | –.36 | .153 |
| task | management | explains the purpose of the lesson.[a] | –.22 | .150 |
| task | instruction | explains everything clearly to me. | –.22 | .150 |
| task | activation | involves me in the lesson. | –.21 | .150 |
| task | activation | encourages me to think for myself. | –.20 | .150 |
| self | climate | ensures that I am relaxed in the classroom. | –.14 | .149 |
| impact | activation | stimulates me to think. | –.10 | .148 |
| impact | activation | ensures that I pay attention. | –.08 | .148 |
| task | management | makes clear when I should have finished an assignment. | .00 | .147 |
| impact | management | applies clear rules. | .04 | .147 |
| task | instruction | ensures that I know the lesson goals. | .18 | .145 |
| task | activation | stimulates my thinking. | .25 | .144 |
| impact | differentiation | connects to what I know or am capable of. | .51 | .141 |
| task | management | ensures that I keep working. | .53 | .141 |
| task | management | ensures that I use my time effectively.[a] | .57 | .141 |
| impact | learning strategies | explains how I should study something. | .63 | .140 |

--- continues next page ---

| stage | domain | teaching practice | *b* | $SE_{(b)}$ |
|---|---|---|---|---|
| impact | differentiation | checks whether I understood the subject matter. [b] | .72 | .140 |
| impact | activation | evokes interest | .76 | .139 |
| impact | differentiation | keeps track of what I know and am capable of.[b] | .78 | .139 |
| impact | learning strategies | teaches me to check my own solutions. | .81 | .139 |
| impact | learning strategies | teaches me to simplify problems. | 1.06 | .137 |
| impact | differentiation | knows what I find difficult. | 1.36 | .135 |
| impact | learning strategies | teaches me to summarize what I have read in my own words. | 1.68 | .134 |

[a] These items had more than five violations for the local independence assumption due to positive increasing correlations in the validation sample.

[b] These items had more than five violations for the local independence assumption due to negative decreasing correlations in the validation sample.

Broadly, the cumulative ordering in Table 3.2 aligns with descriptions of teacher development: It starts with teaching practices that establish a safe learning climate and quality of instruction and ends with teaching practices associated with differentiation and teaching learning strategies. This result confirms our predicted cumulative ordering in complexity. Furthermore, the ordering in Table 3.2 shows considerable within-domain variation, for efficient classroom management in particular. This result suggests that the least complex skills of (more complex) domains may precede the development of the most complex skills of other (less complex) domains. This finding fits with discussions about the limitations of perceiving teacher development in rigid stages, which have continually suggested that descriptions (and measurement) of development in invariant stages is inappropriate and that more flexibility is desirable. By establishing the cumulative ordering at the level of teaching practices, the instrument avoids the requirement of a complete invariant hierarchical ordering in domains. We also note that the ordering includes teaching practices from all six previously identified domains of effective teaching. Our strict procedures for selecting teaching practices thus did not exclude any domain from the

instrument; omitting 27 items describing various teaching practices seemingly did not produce any unacceptable loss of information. In support of this assertion, we computed the correlations of the evaluation scores for teaching skill measured with the original 59 items versus those measured by the 32 selected items. A high correlation would suggest that excluding the 27 items had a minor impact on final evaluations of teaching skill. Indeed, we find that the Pearson product moment correlation between teacher skill scores obtained from the 59- versus 26-item instrument was $r = .99$, with $n = 84$ and $p < .00$.

### 3.4.4 Measurement reliability

To further explore the instrument's properties, we estimated the group-level reliability and SEMs. The group-level reliability (Raju et al., 2006) has similar interpretation to Cronbach's alpha; for students, $\rho_{(\theta\theta')S} = .80$, and for teachers, $\rho_{(\theta\theta')T} = .86$. This result suggests that the instrument reliably discriminates between teachers of different skill. Table 3.3 presents the local SEM estimates associated with the 32 possible response vectors (response vectors with missing values were omitted). The results presented in Table 3.3 suggest increasing measurement precision for skill estimates located more near the center of the measurement scale. For teachers, measurement precision also depend on the number of raters. The Table 3.3 further reveals a ceiling effect for the individual student response vectors. It seems thought, that the discrimination between teachers relies on those 71.1% of the students not rating the teacher as "perfect". This is an issue of concern.

**Table 3.3**

Marginal estimates of the SE as a function of θ for students and teacher

| | students | | | teachers | | | |
|---|---|---|---|---|---|---|---|
| **Raw score** | $f_{(obs.)}$ | $M(\theta)$ | $SE$ | $f_{(obs.)}$ | $n_{(raters)}$ | $M(\theta)$ | $SE$ |
| **1** | 0 | — | — | 0 | — | — | — |
| **2** | 0 | — | — | 0 | — | — | — |
| **3** | 1 | –2.79 | .662 | 0 | — | — | — |
| **4** | 3 | –2.66 | .572 | 0 | — | — | — |
| **5** | 3 | –2.96 | .555 | 0 | — | — | — |
| **6** | 2 | –1.83 | .534 | 0 | — | — | — |

--- continues next page ---

| | students | | | teachers | | | |
|---|---|---|---|---|---|---|---|
| **Raw score** | ***freq. obs.*** | **M(θ)** | ***SE*** | ***freq. obs.*** | $n_{(raters)}$ | **M(θ)** | ***SE*** |
| **7** | 3 | –3.22 | .527 | 0 | — | — | — |
| **8** | 2 | –2.07 | .510 | 0 | — | — | — |
| **9** | 5 | –1.97 | .505 | 0 | — | — | — |
| **10** | 9 | –1.81 | .504 | 0 | — | — | — |
| **11** | 13 | –2.08 | .500 | 0 | — | — | — |
| **12** | 5 | –1.79 | .518 | 0 | — | — | — |
| **13** | 9 | –1.68 | .516 | 0 | — | — | — |
| **14** | 10 | –1.59 | .499 | 0 | — | — | — |
| **15** | 17 | –2.12 | .520 | 0 | — | — | — |
| **16** | 8 | –1.76 | .519 | 1 | 18 | –2.34 | .347 |
| **17** | 15 | –1.50 | .511 | 2 | 28 | –2.12 | .407 |
| **18** | 14 | –1.68 | .509 | 0 | — | — | — |
| **19** | 26 | –1.59 | .514 | 0 | — | — | — |
| **20** | 20 | –1.44 | .521 | 1 | 15 | –1.63 | .380 |
| **21** | 24 | –1.36 | .541 | 1 | 16 | –1.53 | .367 |
| **22** | 44 | –1.29 | .537 | 1 | 2 | –.86 | .737 |
| **23** | 36 | –1.26 | .536 | 0 | — | – | — |
| **24** | 44 | –.95 | .542 | 6 | 97 | –1.02 | .373 |
| **25** | 45 | –.91 | .561 | 7 | 93 | –.81 | .415 |
| **26** | 69 | –.62 | .580 | 10 | 142 | –.59 | .408 |
| **27** | 74 | –.50 | .599 | 14 | 246 | –.36 | .370 |
| **28** | 84 | –.33 | .624 | 9 | 141 | –.14 | .406 |
| **29** | 112 | –.05 | .661 | 9 | 160 | .17 | .381 |
| **30** | 103 | .17 | .733 | 14 | 247 | .56 | .399 |
| **31** | 160 | .56 | .842 | 6 | 93 | 1.19 | .461 |
| **32** | 391 | 1.26 | NA | 3 | 53 | 1.69 | .465 |

### 3.5 Conclusion

Our results confirm the main premise: Effective teaching practices can be ordered cumulatively, from basic to more complex. Broadly, the cumulative ordering observed is in accordance with Fuller's (1969) theory on teacher development, which states that teachers

are first concerned with the self, then with the task, and finally with their impact on student learning. Furthermore, the cumulative ordering mirrors the ordering found on the basis of classroom observations (e.g., van de Grift, et al., 2014; van der Lans, et al., 2017). Thereby, the validation of a cumulative ordering also provides some initial insights in the development of effective teaching practice. These findings represent an important step toward instruments that can provide truly formative feedback. In the future, the instrument developed here could provide an alternative to those in use currently, which can score teachers' current skill but lack the underlying, empirically validated, cumulative ordering required to present objective advice about the next steps to improve.

### 3.5.2 Limitations

We note that the limited sample size of only one school restricts generalization of our findings to other contexts. The results should be viewed in a broader attempt to validate the proposed instrument and its underlying theory. Recently, Maulana, Helms-Lorenz, & Van de Grift (2015) published their findings for a sample of student teachers.

We estimated item fit without consideration of the second teacher level. Our rationale is that in IRT models, there should be a strict separation between the measurement and the structural model. In IRT, and specifically in Rasch models, no interaction between model parts is allowed. In multilevel extensions however, this strict separation is more difficult to attain. The Venn diagram in Figure 3.3 gives the three facets involved and their variances. As Figure 3.3 (next page) shows, the item × student interaction is negligible and, from an IRT perspective, well handled by the model. However, the item × teacher interaction, though small, is not negligible. This violates the assumed strict separation. We present this result to urge the development of multilevel IRT *item* fit tests, which—to our knowledge—are currently not available in IRT software. Current analyses are limited by the unavailability of such fit tests.

**Figure 3.3**

Venn diagram representing the variance decomposition (%) of the facets teacher (t), students nested in teachers (s:t), and item (I) and their interactions. The dashed circle represents the fixed item effect, and the solid lines represent the random teacher (wider circle) and student (inner circle) effects.