Teacher evaluation through observation

van der Lans, Rikkert

Chapter 2

Teacher evaluation on the basis of Classroom Observation

Abstract

This study connects descriptions of effective teaching practices with theory of teacher development to explore an initial understanding how effective teaching develops. The study's main premise is that effective teaching develops cumulatively where more basic teaching practices are required before teachers can develop and use the more complex teaching practices. The sample incorporates teaching practices observed across 878 classrooms. Teaching practices were observed using the International Comparative Analysis of Learning and Teaching (ICALT) observation protocol. Using Rasch Analysis, the study reveals that 31 of 32 effective teaching practices fit cumulative ordering. The ordering also parallels descriptions of teacher development. Together the results indicate that the instrument is a potentially useful tool to describe teachers' development of effective teaching.

Modified version of the article: Van der Lans, R. M., Van de Grift, W., J., C., M., & Van Veen, K. (2017). Developing an instrument for teacher feedback: Using the Rasch model to explore teachers' development of effective teaching strategies and behaviors. *Journal of experimental education* (first online publication). doi: 10.1080/00220973.2016.1268086

2.1 Introduction

Advised and inspired by various reports (e.g., Mourshed, Chijioke, & Barber, 2010), policy makers currently view teacher evaluation and accountability as a primary tactic to improve education. Patrick and Mantzicopolous (2016) provide a comprehensive introduction to these accountability policies. These policies view teachers as accountable for their contribution to students' achievement. Evaluation instruments are used to identify ineffective teachers – i.e. teachers who in comparison to their colleagues contribute little to their students' achievements. Teachers who have been identified as ineffective are given an opportunity to improve. When identified in two or more – depending on the State – consecutive years the teacher should be removed from practice (e.g., National Council on Teacher Quality [NCTQ], 2013). The possibility that teachers may be dismissed on the basis of their achievement data attracts considerable attention due to its extreme personal consequences, yet it includes only a minority of the teacher workforce (Winter & Cowen, 2013). Therefore, in most instances these policies will require ineffective teachers to improve their effectiveness.

However, most evaluation measures are almost exclusively dedicated to the identification of effective teaching practices and as such they provide few information about how to improve effectiveness. The most extreme case are the value-added measures which have been criticized to provide virtually no information about teaching practices used inside the classroom (Darling-Hammond, Amrein-Beardsley, Heartel, & Rothstein, 2012), but also classroom observations of teaching, though clearly providing more information about teaching, also do not completely resolve the underlying problem: i.e., theory of teacher effectiveness has focused on identifying and clustering effective teaching practices, but generally lacks an understanding about how effective teaching develops.

If it is important to observe teaching, then providing teachers with feedback also requires an understanding of teacher development. Several theories of teacher development have been proposed (e.g., Berliner, 2001; Fuller, 1969), but in general, this line of research has unfolded in isolation from research about teacher effectiveness. We seek to combine these streams by turning to the development of effective teaching practice and examining whether effective teaching practice develops cumulatively, in line with what we know about teacher development. If a developmental order in effective teaching practices is acceptable, classroom observation instruments eventually might apply this information to

scaffold feedback that is relevant to the teacher's current stage of development and thereby maximize learning.

In the background theory, we provide a rationale grounded in both the theory on teacher development and prior findings about teacher effectiveness. This synthesis results in a testable hypothesis, predicting stagewise cumulative development in effective teaching practices. The central research question is: Can classroom observations of effective teaching practices be ordered cumulatively? And what does this ordering learn us about the development of effective teaching?

2.2 Background

2.2.1 Teacher development

Teacher development has been described in terms of cumulative phases in expertise or concerns (e.g., Berliner, 2001; Conway & Clark, 2003; Day, Sammons, Stobart, Kingston, & Gu, 2007; Fuller, 1969; Huberman, 1993). These research findings show considerable consistency, despite some points of disagreement, such as about the extent to which teacher development is idiosyncratic, and some differences in research scope, such that studies range from descriptions of effective teaching practices observable in the classroom (e.g., Berliner, 2004) to descriptions of complete life-phases that include factors outside the school (e.g., Day et al., 2007; Huberman, 1993). For this study, we use Fuller's (1969) theory of teacher concerns and incorporate other findings into this framework.

2.2.2 Fuller's theory of teacher concerns

Fuller's (1969) stage theory of teacher concerns was among the first theories of teacher development. It describes teacher development by analyzing trends in teachers' self-reported concerns. Fuller's theory in turn has stimulated two strands of research: one dedicated to describing the development of teaching, and another dedicated to evaluating teacher concerns in the context of innovation and reform (e.g., Richardson & Placier, 2001). This article contributes to the first, in that we seek to describe, evaluate, and measure the development of teaching. In addition, we note that Fuller's initial theory has undergone some changes as the field has developed (Conway & Clark, 2003). Its most recent description entails a relatively simple three-stage model: concerns with the self, concerns with tasks, and concerns with the impact on student learning. Finally, we admit that Fuller's (1969) concerns all pertain to non-behavioral concepts which cannot be observed directly.

Other teacher development theories share this focus alike. For example, Berliner's (2001) phases in teaching expertise mainly involve teacher cognition and information processing. However, Fuller (1969) assumes that teachers' concerns relate to actual behavioral difficulties encountered in the classroom, and Berliner (2001) suggests that teacher cognition defines the limits of teachers' teaching performance. That is, previous theory on teacher development has the auxiliary assumption that differences in teachers' development of concerns and cognitive expertise should result in observable differences in teachers' development of effective teaching practice.

Concerns with the self. Fuller's (1969) first stage, "teachers' concerns with the self," suggest that teachers are initially concerned about their ability to establish respect, trust, and relationships with students (and colleagues). Therefore, Fuller proposes that teachers' development starts with learning how to establish relationships and a constructive learning climate in the classroom. This claim has been corroborated by other research findings. Wubbels and Brekelmans (2005) review two decades of research on interpersonal relationships and found that classroom observations of beginning teachers show more variation in their relationships with students than do those of more experienced teachers. They infer from this result that teacher initial development should focus on relationships. Huberman (1993) reports, on the basis of longitudinal research into pedagogical mastery, that approximately one-third of teachers consider themselves "too close" with students at the beginning of their careers, and another one-third estimates themselves as "too distant." Also, Huberman concludes that beginning teachers should start developing skill in establishing constructive teacher-student relationships. In addition, some teacher observation protocols assign respect and relationships a central position in the development of effective teaching. For example, based on Bowlby's attachment theory the classroom assessment scoring system (CLASS) posits that only in classrooms where students feel safe they will start to learn (Pianta & Hamre, 2009).

Concerns with tasks. The second stage, teachers' concern with tasks, involves concerns about content adequacy, content explanation, and the ability to mobilize resources (Fuller, 1969). Based on these results, Fuller proposes that teacher development proceeds with the development of classroom routines for instruction and management. This claim is corroborated by theories on development in teacher expertise (e.g., Berliner, 2001; Kagan, 1992; Sternberg & Horvath, 1995). These studies generally hypothesize that routines in

management and instruction are prerequisites to move from competent teaching to expert teaching.

Concerns with the impact on student learning. Fuller's (1969) final stage, concerns about the impact on student learning, refers to the teacher's capability to specify objectives for individual students, understand student capacities, and determine how to partial out their own contributions to student difficulties. This view suggests that active teaching methods and differentiation are among the last and most complex teaching practices to develop. Although they are widely recognized as means to promote effective learning (e.g., Hattie, 2009), few studies explore the development of more experienced teachers (for exceptions see: Berliner, 2001; Huberman, 1993). In contrast with the relatively homogeneous development of more elementary stages, the development of more complex teaching practices appears much more varied among teachers, and some teachers never acquire them. Berliner (2001) therefore suggests that deliberate practice, for which formative feedback is key, is required to advance past basic practices.

2.2.3 Teacher effectiveness literature

Several reviews and meta-analyses report on categories of observable teacher behaviors, strategies, or practices – which are here referred to as practices – that contribute to student learning (e.g., Hattie, 2009; Kyriakides, 2013; Marzano, 2003). From these works a vast array of observational instruments have been constructed. Overviews of teaching observation instruments frequently implemented in the U.S. are provided by Patrick and Mantzicolpous (2016), Darling-Hammond (2013), Strong (2011), and Kane et al. (2012). A teaching observation instrument which is currently widely implemented in the Netherlands is the International Comparative Analysis of Learning and Teaching (ICALT) (Van de Grift, 2014). The ICALT refers to these categories with the term "domains" and describes effective teaching by six domains: creating a safe learning climate, efficient classroom management, quality of instruction, student activation, teaching learning strategies, and differentiation. Van de Grift (2014) presents a literature review that detail the six domains, and Maulana, Helms-Lorenz & Van de Grift (2015) detail the connections of these six domains with both the classroom assessment scoring system (CLASS) and the framework for teaching (FFT) teacher observation systems.

2.2.4 Integration of teacher development and teacher effectiveness literature

This study's premise is that observations of effective teaching practices can be related to Fuller's (1969) stages of teacher concerns. Specifically, we hypothesize that teaching practices associated with the domain of a safe learning climate are the least complex, such that they measure and describe the first stage (self) in teacher development. Teaching practices associated with the domains of efficient classroom management and quality of instruction in turn have moderate complexity and together measure and describe the second stage (tasks) in teacher development. Finally, teaching practices associated with the domains of activation, teaching learning strategies, and differentiation are the most complex, so in combination, these practices measure and describe the third stage (impact) in teacher development.

Some previous research offers support for these predictions. Van de Grift, Van der Wal, & Torenbeek (2011) uncover a similar stagewise progression in teaching practices among a sample of primary education teachers. In addition, Kyriakides, Creemers, and Antaniou (2009) report a similar cumulative ordering, using student observations of primary education teachers. Maulana et al. (2015) provide evidence of this ordering using student questionnaire data of beginning secondary education teachers (< 3 years of experience) and Van de Grift, Helms-Lorenz, & Maulana (2014) provide evidence of this ordering using classroom observation of beginning secondary education teachers. Finally, Van der Lans, Van de Grift, Van Veen (2015) also report that student ratings of more experienced secondary education teachers can be ordered cumulatively. This study aims to contribute new evidence of the validity of this cumulative order for evaluation of more experienced secondary education teachers.

2.3 Method

2.3.1 Sample

The sample consisted of 958 teachers whose lessons were observed by trained observers in 119 schools located across the Netherlands. The observations were performed by either peers (53%) or inspectors from the Dutch inspectorate (47%). Teacher experience ranged from student teachers with 0 years of experience to those who had been teaching for 41 years. Of these teachers, 51% were men, and 25.7% held a master's degree. The sample included all education types, from preparatory secondary vocational education to university preparatory education, and students from all grades. The classroom subjects in which the

observations took place were Dutch, history, math, biology, geography, English (as a foreign language), social science, science and physics, economics, French, German, philosophy, arts, drawing and construction, Spanish, Latin, music, and informatics. All observations took place between spring 2010 and summer 2011.

2.3.2 Instrument

The International Comparative Analysis of Learning and Teaching (ICALT) observation instrument includes 32 items that specify observable teaching practices (see the Appendix B). The items refer to six domains—*safe learning climate* which describes the relation between teacher and class; *classroom management* which describes the overall order in the classroom; *clear instruction* which describes the quality explanations of lesson topics and the overall lesson structure, as well as connections among lesson parts; *activation* which mentions various teaching practices that motivate students to think about the topic; *learning strategies* which describes teachers' efforts to teach students how to learn; and *differentiation* which describes whether teachers are sensitive and flexible to meet individual students' learning problems and needs—that together describe the latent variable teaching skill. Observers rated the items on a four point scale (1= "mostly weak"; 2 = "more often weak than strong"; 3 = "more often strong than weak"; 4 = "mostly strong").

2.3.3 Data selection procedures and missing data

Not all classroom observations were completed. We discarded observational forms that counted missing values on more than one-thirds of the items (n = 30). In addition, we discarded classroom observations with missing values on one entire domain (n = 50). After this process, 878 of the original 958 sampled teachers remained. They accounted for 28,096 item responses with 3.38% missing values. We considered these 3.38% missing values to be missing at random.

The 878 classroom observations were randomly divided into a development sample (n = 439) and a validation sample (n = 439). We used the development sample to test the hypothesis of cumulative item ordering and identify items that failed to fit the cumulative ordering. Subsequently, we used the validation sample to cross-validate any evidence of stagewise development in effective teaching practices among the items.

2.3.4 Model

Effective teaching practices should show cumulative, stagewise development. To test this hypothesis, we examined whether the classroom observations of the ICALT fit the three Rasch model assumptions (DeMars, 2010), namely:

1. *Parallel item characteristic curves (ICCs).* This assumption states that descriptions of effective teaching practices discriminate equally among levels of teaching skill.
2. *One-dimensionality.* Descriptions of effective teaching practices can be ascribed to a single latent construct: teaching skill.
3. *Local independence.* The residuals of item pairs are uncorrelated.

We deliberately chose the strict Rasch model instead of the two-parameter item response theory (IRT) model. The only difference between the Rasch model and the two-parameter IRT model is that the latter does not specify the parallel ICC assumption and adds an additional a-parameter that describes the random variation in the steepness (slope) of the ICC's. However, testing whether ICC's are parallel is a prerequisite for evaluating whether cumulative item ordering is plausible (Bond & Fox, 2007).

Note also that we chose to work with the dichotomous Rasch model instead of the polytomous versions of the model. Therefore, the original scoring 1 and 2 are recoded 0 = "insufficient" and the original coding 3 and 4 are recoded 1 = "sufficient". A polytomous model brings in additional complexity which appears considerably confusing to teachers when providing them feedback. Therefore, observers are explicitly trained to adequately distinguish between "insufficient" (1 or 2) and "sufficient" (3 or 4) and observation training procedures require that observers have above 70% inter-rater agreement on the dichotomous "insufficient" or "sufficient". We analyzed whether the dichotomization leads to an unacceptable loss of information. When using the dichotomous Rasch model, the total variance in evaluation outcomes decreases slightly; the range of the polytomous model is 9.58 and the dichotomous model is 8.72. The correlation between the polytomous and dichotomous model is $r(df = 784) = .91$. This evidence gives the impression that the dichotomization does not lead to an unacceptable loss of information.

2.3.5 Data analysis and software

To examine and verify the parallel ICC assumption, we compared the fit of the Rasch model with the fit of the two-parameter IRT model that allows for random ICCs. The

analysis was performed in R using the package ltm (Rizopoulos, 2006). To examine and verify the assumption of one-dimensionality, we applied confirmatory factor analysis (CFA) and in addition we report on the scree plot of the exploratory factor analysis (EFA). The analysis was performed in Mplus (Muthén & Muthén, 1998–2012). The CFA model constrains factor loadings to be 1.00 and residual correlations to be zero. Furthermore, the variance of the factor is standardized to 1.00. The estimation algorithm we used is "WLSMV"; the parameterization is "Theta". The EFA explored a one and two factor solution using a geomin oblique rotation with the estimation algorithm "WLSMV". To examine and verify local independence, we applied two tests: (1) Ponocny's (2001) T_1 and T_{1m} tests, (2) and Chen and Thissen's (1997) LD-χ^2 test. Ponocny's tests were estimated in R using the eRm package (Mair & Hatzinger, 2007). The Chen and Thissen LD-χ^2 test is estimated using IRT-PRO (Cai, Thissen, & Du Toit, 2005–2013). Finally, the cumulative item ordering is estimated using a multilevel Rasch model where teachers are nested in schools. The item parameters were estimated using the R package lme4 (Bates, Maechler, Bolker & Walker, 2014).

2.4 Results

We first report the results of our empirical analysis of cumulative ordering, including the fit of the three Rasch model assumptions, and then present the evaluation instrument, its cumulative ordering, and a comparison with Fuller's (1969) stage theory.

2.4.1 Development sample

Parallel ICC. The Rasch model and two-parameter IRT model are nested models that differ only in that the latter allows for random item characteristic curves (ICC). Rizopoulos (2006) suggested comparing the fit of both models using the $\Delta\chi^2$ test. If the $\Delta\chi^2$ test is insignificant, the ICC's are approximately parallel. The results indicate that the Rasch model has slightly worse fit than the two-parameter IRT model ($\Delta\chi^2 = 45.14$, $df = 31$, $p < .05$). Closer inspection of the random slope parameters the (a-parameters) reveals that item slopes varied from 1.30 (SE = .23) < a < 2.46 (SE = .42). These slopes do not statistically deviate (±1.96*SE) from the average slope ($M(a)$ = 1.75). The only exception is item 22, "explains the lesson objectives at the start of the lesson," for which the ICC slope (a) = 1.05, SE = .18. When deleting item 22, the Rasch model and two-parameter IRT

model have identical fit ($\Delta\chi^2 = 36.54$, $df = 30$, $p = .19$). Therefore, all items other than item 22 exhibited approximately parallel ICC.

One-dimensionality. The assumption of one-dimensionality is difficult to (dis)confirm. Despite the fact that many tests have been proposed to evaluate one-dimensionality (e.g., Haberman, 2008; Stout, 1990; Timmerman, Lorenzo-Seva, & Ceulemans, in press), there is not much consensus about any statistical approach. In addition, there is considerable discussion about the best criteria with which to evaluate the goodness of fit of statistical models. Some propose to use exact-tests which can reject the null-hypothesis of one-dimensionality, such as the χ^2-statistic in confirmatory factor analysis (CFA) (Kline, 2011) or Kelley's regression formula (Haberman, 2008). Others point out that an exact-test of one-dimensionality is overly strict and often rejects the null-hypothesis even when the data can be appropriately described using one dimension (e.g., DeMars, 2010; Steiger, 2007; Stout 1990). They therefore propose to use "approximate fit" indices such as the root mean square error of approximation (RMSEA) or to use an approach based on some type of ratio between eigenvalues.

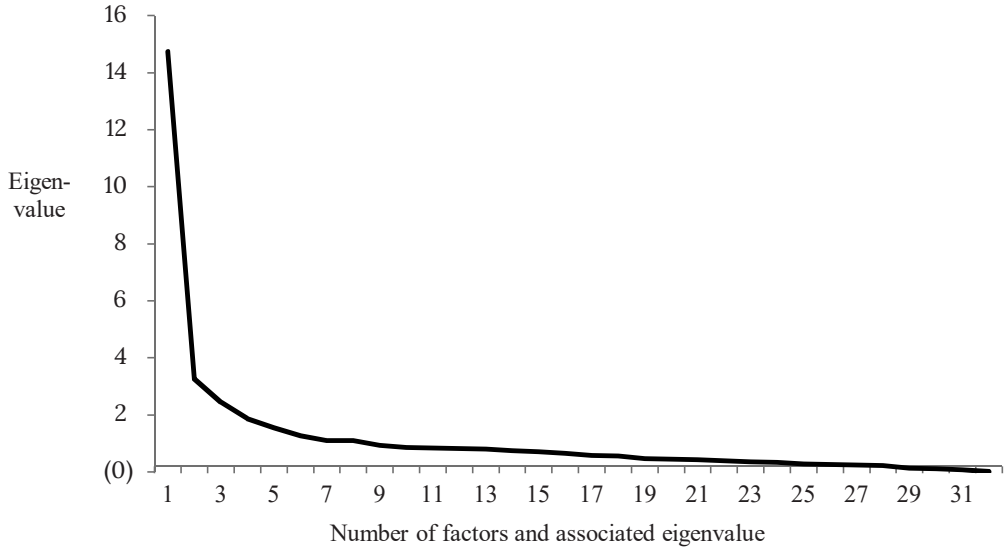
We therefore apply confirmatory factor analysis to explore whether the one-dimensional solution provides a reasonable description of the data. To evaluate model fit we rely on approximate fit indices, in specific the root mean square error of approximation (RMSEA), the Comparative Fit Index (CFI), and the Tucker-Lewis Index (TLI). We apply the criteria RMSEA < .05, CFI < .95, and TLI < .95, as has been recommended by Hu and Bentler (1999). As input, we used the tetrachoric item correlations instead of Pearson phi correlations, as DeMars (2010) recommends.

The results of the CFA for the one-factor model present a mixed picture (χ^2 (496, $n = 439$) = 950.68, $p = .00$; CFI = .93, TLI = .93, RMSEA = .046 [90% CI = .041, .050]). While the RMSEA indicates close fit, the CFI and TLI are below the threshold of .95. To depict this result, we added the scree plot (Figure 2.1) produced by an EFA. The plot clearly shows one dominant factor: The eigenvalue of the first factor is 14.54 and nearly five times greater than the eigenvalue of the second factor, at 3.06. A brief examination of the EFA factor loadings shows that they are all within the range of .61 to .81. The only exception is item 22, "explains the lesson objectives at the start of the lesson," which had a factor loading of .45. Removing item 22 slightly improved fit (χ^2 (465, $n = 439$) = 841.49, $p = .00$; CFI = .94, TLI = .94, RMSEA = .043 [90% CI = .038, .048]), but CFI and TLI remained below .95. Further improvement of model fit requires freeing residual correlations

between item pairs. This indicates that the observed deviations from one-dimensionality are due to violations of local independence which will be investigated next.

Figure 2.1

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



Note. The y-axis shows the eigenvalues, and the x-axis shows the number of factors.

Local independence. We investigated local independence using the nonparametric T_1 and T_{1m} statistics (Ponocny, 2001) and the LD-χ^2 test (Chen & Thissen, 1997). Given the current stage of the instrument and theory development, we considered false positives (i.e., retaining items that violate local independence) less severe than false negatives (i.e., removing items that do not violate local independence). Concurrently, our concerns are for inflation of the alpha level. We describe and report the results of Ponocny's (2001) T-tests first, then reevaluate the results with the Chen and Thissen (1997) LD-χ^2 test.

Ponocny's T_1 and T_{1m}. Rasch (1960) originally proposed but never completed a nonparametric test to assess model fit. Ponocny (2001) based his family of *T*-statistics on Rasch's original intentions, using the raw sum scores for the items and persons. From the observed sum scores, this test generates alternative matrices with identical sum scores, then tests whether these alternative matrices all fit the Rasch model. The test of local independence uses the number of extreme scoring patterns: {11} or {00} and tests whether

the number of extreme scoring patterns on two items *j* and *k* is higher (in case of T_1) or lower (in case of T_{1m}) than the number of extreme scoring patterns expected by the Rasch model (see Koller & Hatzinger, 2013; Ponocny, 2001). With an instrument of 32 items, the tests evaluate violations of local independence for 496 item pairs. With so many tests, some violations may occur simply due to chance (e.g., Koller and Hatzinger, 2013; Ponocny, 2014, personal communication). Koller and Hatzinger (2013) therefore propose correcting for chance inflation by dividing alpha by the number of item pairs tested. With this correction, the alpha level becomes (.05/496) = .0001.

Ponocny's (2001) T_{1m} test diagnoses two item pairs, showing decreasing residual correlations. Negative residual correlations indicate that the teaching practices described by two items each have an additional characteristic, other than teaching skill due to which item scores are less similar than can be explained by teaching skill alone. Two item pairs shared such a negative residual correlation: item 1, "shows respect for students in behavior and language" (domain of creating a safe learning climate) with item 22, "explains the lesson objectives at the start of the lesson" (domain of student activation), and item 5, "ensures that the lesson runs smoothly" (domain of efficient classroom management) with item 24, "offers weak students additional learning and instruction time" (domain of differentiation).

The test results indicate that the teaching practices of presenting lesson goals and showing respect for students share a negative dependency that is independent of teaching skill. Note that item 22 misfits all model assumptions. Due to this we give no further substantial interpretation to this item pair. The other item pair suggests that teaching practices focused on 'offering additional time to weak students' and those required to 'ensure smooth running lessons' share a negative dependency. We speculate that it might be that some teachers differentiate between students, but their chosen method of doing so negatively affects their classroom management.

Next, the T_1 tests reveal any positive increasing residual correlations. Positive residual correlations indicate that thee teaching practices described by two items share an additional characteristic, other than teaching skill, due to which item scores are more similar than can be explained by teaching skill alone. This test diagnosed six item pairs (we list their domains in parentheses): item 3, "supports student self-confidence" (safe learning climate) with item 17, "boosts the self-confidence of weak students" (student activation); item 21, "provides interactive instruction" (student activation) with item 31, "encourages students to think critically" (teaching learning strategies); item 22, "explains the lesson

objectives at the start of the lesson" (student activation) with item 23, "checks whether the lesson objectives have been achieved" (differentiation); item 24, "offers weak students additional learning and instruction time" (differentiation) with item 25, "adapts processing of subject matter to student differences" (differentiation); item 27, "teaches students how to simplify complex problems" (teaching learning strategies) with item 32, "asks students to reflect on approach strategies" (teaching learning strategies); and finally, item 28, "encourages the use of checking activities" (teaching learning strategies) with item 29, "teaches students to check solutions" (teaching learning strategies).

Some of these results may reflect similarities in the item phrasing, such as when items 3 and 17 refer to "supports self-confidence" and "boosts self-confidence." Most of the diagnosed pairs share domain membership though, such that items 24 and 25 both represent differentiation, and items 27 and 32, as well as 28 and 29, represent the teaching learning strategies domain. Particularly for these more complex domains, classroom observers might experience more difficulty clearly understanding and discriminating among distinct teaching practices. An exception is item pair 21-31. Interactive instruction frequently involves interactively posing questions. This residual correlation might indicate that some observers have come to view 'interactive instruction' as an alternative phrasing of 'encouraging critical thinking'.

Finally, we used the LD-χ^2 test proposed by Chen and Thissen (1997). It approaches local independence, as is the case in which the observed frequencies of the responses 0 and 1 on two items j and k do not deviate from their expected frequencies, based on the trace line. Deviations between observed and expected frequencies then can be tested against a chi-square distribution with one degree of freedom. To correct for chance, we diagnosed item pairs for which $\chi^2 > 15.14$, because the test df is equal to 1, which results in an alpha value of .0001. The LD-test diagnosed four item pairs: items 5–24, items 24–25, items 25–26, and items 28–29. With the exception of items 25–26 (both in the differentiation domain), these item pairs also were diagnosed previously by Ponocny's T_1 and T_{1m}. Descriptions of the items appear in Table 2.1.

2.4.2 Summary of main findings for the development sample

In the development sample, only item 22, "explains the lesson objectives at the start of the lesson," exhibited misfit with the cumulative stagewise pattern. It has a deviating ICC and a low factor loading, and one test (Ponocny's T_{1m} and T_1) confirmed that it violates local

independence. We note further that item 22 previously has been shown to violate model assumptions (see for example: Van de Grift, Helms-Lorenz, & Maulana, 2014).

2.4.3 Cross-validation

In the validation sample (n = 439), we readdressed all three assumptions. The chi-square difference test between the one- and two-parameter models indicated some violations of the parallel item characteristic curves (ICC) assumption ($\Delta\chi^2$ = 58.49, df = 31, p = .02). Exclusion of item 22, which again had the lowest discrimination parameter, improved model fit but insufficiently ($\Delta\chi^2$ = 54.61, df = 30, p = .04). An additional examination of the discrimination parameters indicated that item 10, "gives feedback to students," also deviated considerably. Its discrimination parameter (a = 3.03, SE = .052) was almost twice as steep as the average discrimination parameter ($M(a)$ = 1.67). After we deleted item 10, the remaining 30 items were found to have approximately parallel ICC ($\Delta\chi^2$ = 41.83, df = 29, p = .06).

We reassessed the assumption of one-dimensionality using a CFA. The model fit again is mixed with RMSEA below the .05 threshold, but CFI and TLI above the threshold (χ^2 (465, n = 439) = 906.69, p = .00; CFI = .94, TLI = .94; RMSEA = .047 [90% CI = .042 - .051]). The scree plot again showed one dominant factor: The first eigenvalue was 14.80, whereas the second was 3.20.

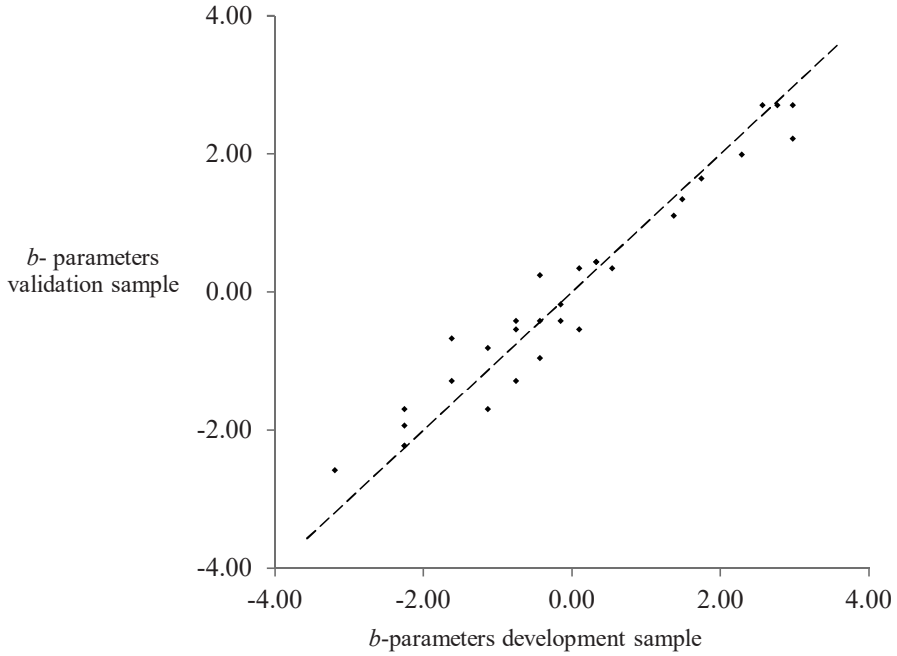
Finally, we reassessed local independence. We report findings that replicate those from the development sample (the complete results are available on request). Ponocny's T_{1m} test again diagnosed two item pairs that violated local independence. The results validated the negative residual correlations between the items in the domain of efficient classroom management and in the domain of differentiation, though the specific item pairs differed. The Ponocny's T_1 test also diagnosed six item pairs, most indicating again positive residual correlations between items in the domains of differentiation and of teaching learning strategies. Finally, the LD-χ^2 test did not diagnose item pairs not already diagnosed by Ponocny's tests.

This cross-validation accordingly confirmed that all items except item 22 fit the invariant cumulative and one-dimensional ordering. We recommend that the item should be discarded from the instrument when a Rasch analysis is applied. Tests for local independence consistently diagnosed some underlying patterns that might help clarify what creates multidimensionality in the current measures of teaching skill. In particular, the

negative residual correlations between effective teaching practices in the domain of efficient classroom management and those in the domain of differentiation request further exploration.

Figure 2.2

The goodness-of-fit (GoF) plot.



Notes. The 31 dots represent the 31 items. 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.

The 31 items show an invariant and cumulative ordering in terms of effective teaching practices. In support of this assertion, we further examined the invariance between samples. Figure 2.2 presents the goodness-of-fit (GoF) plot, in which dots indicate each item, and the dashed line reflects perfect invariance between samples (i.e., zero-difference score). The deviations from the dashed line indicate deviations from invariance. We used the Andersen's (1973) LR test to examine whether the two samples showed such deviations, but the results indicated no such deviation ($\Delta\chi^2 = 41.86$, $df = 30$, $p = .07$).

2.4.4 Comparison of cumulative ordering with Fuller's theory

Table 2.1 presents the obtained cumulative ordering of teaching development. More complex teaching practices are denoted by higher *b*-parameters. Broadly, the cumulative ordering obtained from the data is in line with Fuller's (1969) previous descriptions of teacher development, in which concern for the self precedes concern for the task, and concern for the task precedes concern for the impact on student learning. We therefore propose that this cumulative ordering represents teacher development and can be applied to provide teachers with feedback about promising directions for their further training and professional development.

Table 2.1

Final cumulative ordering in effective teaching practices

stage	domain	teaching practice	*b*	$SE_{(b)}$
self	climate	shows respect for students in behavior and language	− 3.19	.272
self	climate	creates a relaxed atmosphere	− 1.53	.177
self	climate	supports student self-confidence	− 1.43	.174
task	management	ensures effective class management	− 1.23	.169
self	climate	ensures mutual respect	− 1.15	.166
task	management	ensures that the lesson runs smoothly	− 1.06	.164
task	instruction	explains the subject matter clearly	− 1.00	.163
task	instruction	gives feedback to students	− .96	.162
task	instruction	clearly explains teaching tools and tasks	− .91	.161
task	management	checks during processing whether students are carrying out tasks properly	− .80	.159
task	instruction	gives well-structured lessons	− .72	.156
task	instruction	involves all students in the lesson	− .56	.153
task	management	uses learning time efficiently	− .51	.153
task	instruction	encourages students to do their best	− .41	.151
task	instruction	checks during instruction whether students have understood the subject matter	− 28	.149

--- continues next page ---

stage	domain	teaching practice	*b*	$SE_{(b)}$
impact	activation	asks questions that encourage students to think	− .05	.147
impact	activation	uses teaching methods that activate students	.18	.144
impact	activation	encourages students to reflect on solutions	.22	.144
impact	activation	provides interactive instruction	.34	.142
impact	activation	boosts the self-confidence of weak students	.35	.143
impact	learning strategies	encourages students to think critically	.67	.140
impact	activation	has students think out loud	.68	.141
impact	learning strategies	encourages students to apply what they have learned	.81	.141
impact	learning strategies	teaches students how to simplify complex problems	.99	.139
impact	learning strategies	encourages the use of checking activities	1.42	.140
impact	differentiation	checks whether the lesson objectives have been achieved	1.49	.139
impact	learning strategies	teaches students to check solutions	1.56	.140
impact	learning strategies	asks students to reflect on approach strategies	1.71	.139
impact	differentiation	adapts processing of subject matter to student differences	2.15	.140
impact	differentiation	offers weak students additional learning and instruction time	2.43	.142
impact	differentiation	adapts instruction to relevant student differences	2.85	.145

Note. Fuller stage, effectiveness domain, and complexity of the teaching practices (b).

The results in Table 2.1 also show that the most complex practices in less complex domains surpass the least complex practices of more complex domains. This pattern suggests that teachers do not develop all the skills in one domain first, before proceeding to the next domain. Rather, the transition from one stage to the next is gradual. Some

domains, such as efficient classroom management and quality of instruction, appear almost equally complex, which suggests that they might develop simultaneously.

2.5 Conclusion

Current educational policies assign teacher evaluation a central position in their efforts to improve education. A consensus holds that classroom observations are most appropriate for teacher evaluations that aim to stimulate further professional development. However, to provide teachers with feedback about how to improve their effectiveness, current knowledge about effective teaching needs to be complemented with an understanding how effective teaching develops. On the basis of Fuller's (1969) theory of stages in teacher concerns, we hypothesized that observations of effective teaching practices show invariant cumulative ordering. Broadly, the study results confirm that 31 of the original 32 effective teaching practices exhibit a cumulative ordering. Also, the ordering strongly parallels Fuller's (1969) stages. We therefore suggest that this ordering describes a stagewise development of effective teaching practices. This development starts by developing practices to achieve a safe learning climate, then proceed to develop teaching practices directed at an efficient classroom management and quality in instruction. If skills in these domains are sufficiently mastered, teachers start developing practices in domains related to activating teaching methods, teaching learning strategies, and differentiating and adapting lesson content to meet particular student needs. Together we conclude that the instrument is a potentially useful tool to describe and evaluate teachers' development of effective teaching.

2.5.1 Limitations

The sample included 958 teachers working in 119 schools. Technically, the data should be considered nested, with teachers nested in schools. While the parameters are estimated using multilevel Rasch model techniques, the assumptions tests have not been corrected for the multilevel structure. We checked whether the item parameters estimated by the assumption tests differed from the parameters estimated in the multilevel Rasch model, to verify validity of the assumption tests. No large deviations were found between them. However, the standard errors of item parameters were larger in the multilevel Rasch model, compared to the standard errors estimated by the assumption tests. This implies that by not correcting for the nested data structure in our assumption tests we plausibly are overly strict

and wrongly removed items that actually did fit. We note that assumption tests to evaluate fit of Rasch models in nested datasets are still at a developmental phase (e.g., de Boeck et al., 2011; Fox, 2010) and have not yet been incorporated in standard software packages. We consider our approach as the best option currently available.

Another limitation concerns the stability with which teaching observation instruments can classify teachers. Patrick and Mantzicopolous (2016) show the considerable fluctuations in observed teaching practice across lessons. Their results would suggest that the identified teacher stage of development may change from one day to another. As a consequence, the advice for improvement will change. We are currently exploring whether multiplying the number of observers and lessons may improve the stability (Van der Lans, Van de Grift, Van Veen & Fokkens-Bruinsma, 2016).

The results in support of the assumption of one-dimensionality are mixed. An explanation can be found in correlations between item residuals. Using Ponocny's (2001) non-parametric *T*-tests, we have diagnosed several item pairs as potential violators of local independence. The results indicate negative residual correlations between items describing teaching practices in the domains of efficient classroom management and differentiation. More precise the number of lesson observations where items in the domain efficient classroom management is scored "insufficient" and items in the domain differentiation are scored "sufficient" is slightly higher than would be predicted by the model. We speculate that it might be that some teachers differentiate between students, but their chosen method of doing so negatively affects their classroom management. Another possibility is that some observers came to see differentiation as providing freedom to students. For example, in Dutch mathematics classes some teachers start their lessons with instruction and explanation after which they write down assignments on the blackboard. They announce that during the remaining time of the lesson students can work in their own pace on the assignments. Such classes can become considerably noisy and unorganized. However, some observers might have wrongly interpreted "working in their own pace" as a teaching practice to differentiate between students. The results of Ponocny's T_1 test indicate that the items in the last two domains share positive residual correlations. Here, the number of lesson observations reporting both items as "sufficient" (teaching learning strategies) or both as "insufficient" (differentiation) is greater than expected by the model. Due to this the items are estimated as more similar in complexity than they actually are. We speculate that classroom observers might experience difficulties understanding these more complex

practices and may not feel confident or lack knowledge about how to discriminate among them.

In conclusion, the residual correlations tend to 'break' the one-dimensional ordering into two factors. Items describing more complex teaching practices tend to cluster together, while also pushing away items in the domain efficient classroom management. Further research is needed as to what are plausible explanations for this.