Developing an Instrument for Teacher Feedback: Using the Rasch Model to Explore Teachers' Development of Effective Teaching Strategies and Behaviors

Rikkert M. van der Lans, Wim J. C. M. van de Grift & K. van Veen


To link to this article: https://doi.org/10.1080/00220973.2016.1268086

© 2018 The Author(s). Published with license by Taylor & Francis Group, LLC
Rikkert M. van der Lans, Wim J. C. M. van de Grift, and K. van Veen.

Published online: 06 Jan 2017.

Article views: 952

View related articles

View Crossmark data

Citing articles: 2 View citing articles
ABSTRACT
This study connects descriptions of effective teaching with descriptions of teacher development to advance an initial understanding of how effective teaching may develop. The study’s main premise is that descriptions of effective teaching develop cumulatively where more basic teaching strategies and behaviors are required before teachers may advance to more complex teaching behaviors. The sample incorporates teaching behaviors observed across 878 classrooms. Teaching behaviors 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 behaviors 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.

KEYWORDS
Rasch model; teacher development; teacher effectiveness; teacher evaluation; teaching quality

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. In their recent contribution to this journal, Patrick and Mantzicopoulos (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 identified are given an opportunity to improve. When identified in 2 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 to dismiss teachers on the basis of achievement data attracts considerable attention due to its extreme personal consequences, yet it includes only a minority of the teacher workforce (Winter & Cowen, 2014). Thus, 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 little 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 behavior (Darling-Hammond, Amrein-Beardsley, Heartel, & Rothstein, 2012); but also classroom observations of teaching, though clearly providing more information about teaching behavior, also do not completely resolve the underlying problem. That is, theory of teacher effectiveness has focused on identifying and clustering effective teaching behaviors but generally lacks an understanding of how effective teaching develops.

If it is important to observe effective teaching behaviors, 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 behaviors and examining whether effective teaching behaviors develop cumulatively, in line with what we know about teacher development. If a developmental ordering in effective teaching behaviors 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 behaviors. The central research question is, Can classroom observations of effective teaching strategies and behaviors be ordered cumulatively, and what does this ordering teach us about the development of effective teaching?

**Background theory**

**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 teacher behaviors 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.

**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 nonbehavioral concepts, which cannot be observed directly. Other teacher development theories share this focus. 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 involves the auxiliary assumption that differences in teachers’ development of concerns and cognitive expertise should result in observable differences in teachers’ development of teaching behaviors and strategies.

**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 find 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 will they 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 how to determine their own contributions to student difficulties. This view suggests that active teaching methods and differentiation are among the last and most complex teaching behaviors 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 behaviors 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 behaviors.

**Teacher effectiveness literature**
Several reviews and meta-analyses report on categories of observable teacher behaviors 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. For an overview of teaching observation instruments frequently implemented in the United States, we refer to Patrick and Mantzicopoulos (2016) and Kane et al. (2012). A teaching observation instrument that 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. The six domains are creating a safe learning climate, efficient classroom management, quality of instruction, activating teaching methods, teaching learning strategies, and differentiation. van de Grift (2014) presents a literature review that details the six domains, and Maulana, Helms-Lorenz, & van de Grift (2015) detail the connections of these six domains with both the CLASS (Pianta & Hamre, 2009) and the framework for teaching (FFT) (Danielson, 2013) teacher observation systems. A brief description of the six domains is in the Appendix.

**Integration of teacher development and teacher effectiveness literature**
The study’s premise is that observations of effective teaching behaviors can be related to Fuller’s (1969) stages of teacher concerns. Specifically, we hypothesize that teaching behaviors in 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 behaviors in 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 behaviors in the domains of activating teaching
methods, learning strategies, and differentiation are the most complex; so in combination, these behaviors measure and describe the third stage (impact) in teacher development.

Some previous research offers support for these predictions. van de Grift, van de Wal, and Torenbeek (2011) uncover a similar stagewise progression in teaching behaviors among a sample of primary education teachers, and van de Grift, Helms-Lorenz, & Maulana (2014) provide evidence of this ordering using classroom observation of beginning secondary education teachers. In addition, Kyriakides, Creemers, and Antaniou (2009) report a similar cumulative ordering, using student observations of primary education teachers. And van der Lans, van de Grift, and van Veen (2015) also found that student ratings of secondary education teachers can be ordered cumulatively. These previous works concentrated on primary education teachers and beginning teachers. Teachers and teaching in secondary education differ in important respects from primary education, so this study will examine the developmental stages exhibited by experienced secondary education teachers.

**Method**

**Sample**

The sample consisted of 958 teachers whose lessons were observed by trained observers in 119 schools across the Netherlands. This sample included observations performed by inspectors from the Dutch inspectorate (47%) and the lesson observations gathered within a university research project. 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. The observations were performed by either peers or inspectors from the Dutch inspectorate. All observations took place between spring 2010 and summer 2011.

**Instrument**

The International Comparative Analysis of Learning and Teaching (ICALT) observation instrument includes 32 items that specify observable teaching behaviors (see the Appendix). 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 overall lesson structure and the connections among lesson parts; activating teaching methods, which mentions various teaching strategies 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 to and flexible in attempting to meet individual students’ learning problems and needs. Items referring to the six domains together describe the latent variable teaching skill. Observers rated the items on a 4-point scale (1 = mostly weak; 2 = more often weak than strong; 3 = more often strong than weak; 4 = mostly strong).

**Data selection procedures and missing data**

Not all classroom observations were completed. We discarded observational forms that counted missing values on more than one-third 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 behaviors among the items.

**Model**

Effective teaching behaviors 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):

1. **Parallel item characteristic curves (ICCs).** Teaching behavior discriminates equally among levels of teaching skill.
2. **One-dimensionality.** Observations of teaching behaviors can be ascribed to a single latent construct—that is, 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 ICCs. However, testing whether ICCs 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.” This choice is made, because a polytomous model brings in additional complexity that appears considerably confusing to teachers when providing them feedback. In the Dutch project 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% interrater agreement on the dichotomous “insufficient” or “sufficient.” Nevertheless, 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 suggests that the dichotomization does not lead to a completely unacceptable loss of information.

**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 0. Furthermore, the variance of the factor is standardized to 1.00. Estimation algorithm used is “WLSMV”; 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 types of tests: (1) Ponocny’s (2001) non-parametric $T_1$ and $T_{1m}$ tests, (2) and Chen and Thissen’s (1997) LD-$\chi^2$ test. Ponocny’s tests were estimated in R using the eRm package (Mair & Hatzinger, 2007). The Chen and Thissen LD-$\chi^2$ 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, et al. 2014).

**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.
**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) suggests 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 ($\pm 1.96 SE$) 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 no 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 $\chi^2$-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 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 ($\chi^2 (496, N = 439) = 950.68, p = .00; CFI = .93, TLI = .93, RMSEA = .046 [90% CI = .041, .050]$). While RMSEA indicates close fit, the CFI and TLI are below the threshold of .95. To depict this result, we added the scree plot (Figure 1) produced by an EFA. The plot clearly shows one dominant factor: The eigenvalue of the first factor is 14.52 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 teaching act 22, “explains the lesson objectives at the start of the lesson,” which had a factor loading of .45. Removing item 22 slightly improved fit ($\chi^2 (465, N = 439) = 841.49, p = .00; CFI = .94, TLI = .94, RMSEA = .043 [90% CI = .038, .048]$), 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.

**Local independence**

We investigated local independence using the nonparametric $T_1$ and $T_{1m}$ statistics (Ponocny, 2001) and the LD-$\chi^2$ 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-$\chi^2$ test.
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 & Hatzinger, 2013; Ponocny, 2014, personal communication). Koller & 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 $0.05/496 = 0.0001$.

Ponocny’s (2001) $T_{1m}$ test diagnoses two item pairs, showing decreasing residual correlations. Negative residual correlations indicate that the two behaviors describe an additional characteristic other than teaching skill owing 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 behaviors 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 behaviors that “offer weak students additional time” 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 two teaching behaviors describe 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

\[ \text{Figure 1. Scree plot of the exploratory factor analysis using the tetrachoric correlations. Note. The y-axis shows eigenvalues, and the x-axis shows the number of factors.} \]
learning climate), with item 17, “boosts the self-confidence of weak students” (activating teaching methods); item 21, “provides interactive instruction” (activating teaching methods), with item 31, “encourages students to think critically” (teaching learning strategies); item 22, “explains the lesson objectives at the start of the lesson” (activating teaching methods), 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 and items 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 behaviors. 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-χ² 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 χ² > 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 and 24, items 24 and 25, items 25 and 26, and items 28 and 29. With the exception of items 25 and 26 (both in the differentiation domain), these item pairs also were diagnosed previously by Ponocny’s T₁ and T₁m.

Descriptions of the items appear in Table 1.

**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 T₁ and T₁) confirmed that it violates local independence. We note further that item 22 previously has been shown to violate model assumptions (see e.g., van de Grift, Helms-Lorenz, & Maulana, 2014).

**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 (Δχ² = 58.49, df = 31, p = .02). Exclusion of item 22, which again had the lowest discrimination parameter, improved model fit but insufficiently (Δχ² = 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 (Δχ² = 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 with CFI and TLI above the threshold (χ² (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-$\chi^2$ 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 teaching behaviors in the domain of efficient classroom management and those in the domain of differentiation request further exploration.

The 31 items show an invariant and cumulative ordering in terms of teaching behaviors. In support of this assertion, we further examined the invariance between samples. Figure 2 presents the goodness-of-fit (GoF) plot, in which dots indicate each item (cf. the omitted item 22) 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 ($\Delta \chi^2 = 41.86$, $df = 30$, $p = .07$).

**Comparison of cumulative ordering with Fuller’s theory**

Table 1 presents the obtained cumulative ordering of teaching development. More complex teaching behaviors 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 suggest 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.
Furthermore, the results show that the most complex behaviors in less complex domains surpass the least complex behaviors 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.

Conclusion and discussion

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 of how effective teaching develops. On the basis of Fuller’s (1969) theory of stages in teacher concerns, we hypothesized that observations of effective teaching behaviors and strategies show invariant cumulative ordering. Broadly, the study results confirm that 31 of the original 32 effective teaching behaviors 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 behaviors. This development starts by developing behaviors to achieve a safe learning climate, then proceeds to develop behaviors for efficient classroom management and quality in instruction. If skills in these domains are sufficiently mastered, teachers start developing behaviors in domains related to activating teaching methods and 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.

**Limitations**

The sample included 958 teachers working in 119 schools. Technically, the data should be considered nested, with teachers nested in schools. The items evaluated, however, are neither nested in teachers nor in schools (e.g., Doran, Bates, Blies, & Dowling, 2007), and it can be shown that item b-parameters are relatively unaffected by the nested data structure. Yet their standard errors are affected and currently most likely biased downward. This implies that by not correcting for the nested data structure our assumption tests plausibly are overly strict. Assumption tests to evaluate fit of Rasch models in nested data sets are still at a developmental phase (e.g., de Boeck et al., 2011) 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 Mantziopoulos (2016) show the considerable fluctuations in observed effective teaching behaviors 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 2016).

**Concerns for model fit and possibilities for improvement**

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 behaviors in the domains of efficient classroom management and differentiation. This implies that items in these domains receive too few equal responses (i.e. {00} or {11}). 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 strategy to differentiate between students. The results of Ponocny’s T1 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” (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 difficulty in understanding these more complex behaviors 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 strategies 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.

**Expected impact on student learning**

The developed instrument is grounded in theory about teacher development and in literature about teaching effectiveness. This study aims to study the developmental progression, but an important
question is what can be expected when teachers succeed to develop more skill. Are teachers succeeding to implement more-complex teaching strategies and more-effective teaching behaviors? In one empirical study (van de Grift & Lam, 1998), the predictive validity of the instrument has been evaluated. This study found a significant effect on student achievement in primary education. Furthermore, we note that the here developed instrument shows much overlap with other instruments currently in use, including the Classroom Assessment Scoring System (CLASS) and the Framework for Teaching (FFT). These instruments have been found to be predictive of student achievement gains (Kane et al., 2012).

**Advantages of the instrument for research**

The insights that the ICALT instrument can provide about the development of effective teaching may prove to be a useful addition to the currently available research instruments. An area of research where the instrument may prove useful involves research examining the effectiveness of new teaching methods and new instructional materials or other interventions in the classroom. An example of research in this area that was recently published in this journal is Cromley et al. (2013). Cromley et al. developed and implemented a classroom instruction method to improve student diagram learning. A potential confounding variable in this study design is teachers’ ability to work with the method. When using the ICALT instrument it is possible to select those specific teachers who have the required skill to work with the new methods and materials—for example, selection may concern teachers who demonstrate skill in classroom management and instruction. In addition, the instrument may be used to verify whether teachers’ teaching skill in the experimental and the control group is comparable. A second area of research in which the instrument may prove useful is teachers’ professional development. Many mostly small teacher training initiatives have been taken to improve and stimulate development of teaching effectiveness. However, most initiatives find only small effects (e.g., Hill, Besiegel, & Jacob, 2013). The ICALT instrument may inform professional development trainers about teachers’ “zone of proximal development”—that is, it seems likely that teachers will learn most effectively those skills that are just above their current skill level. The instrument may inform about which specific teaching behaviors and strategies to address. Again researchers investigating the effectiveness of their training program may want to select teachers for whom the training is expected to be most effective.

**Practical implications, use, and applicability**

The main goal of the instrument is to provide teachers with a diagnostic indication about current proficiency and the most promising directions for further teacher training and self-reflection. We believe the instrument offers great potential to contribute to the provision of formative feedback in teacher evaluations. From an evaluation perspective, the main advantage of a cumulative item ordering that reflects complexity levels is that the teacher’s evaluation score can be used to scaffold feedback to the appropriate level of competence. Specifically, teaching behaviors whose complexity is near the teacher’s teaching skill are most relevant for further training and self-reflection; teaching behaviors whose complexity exceeds the teacher’s teaching skill by far will be too complex to develop through training, and teaching behaviors that are well below the teachers’ teaching skill likely already have been acquired by that teacher. So, using this cumulative ordering, individual teachers can receive a diagnosis about the most plausible ways to proceed with their professional development. Note that these diagnoses are meant to provide initial starting points for further training and learning. That is, items descriptions give an understanding where to look for improvement. For example, the item “ensures mutual respect” functions as an umbrella under which a range of behaviors can be specified about how to achieve mutual respect. Based on this diagnosis, a teacher could be advised to read theory about teacher-student relationships (e.g., Wubbels & Brekelmans, 2005; Pianta & Hamre, 2009) or to discuss with colleagues what possible strategies to follow in order to improve “mutual respect.” She/he could choose to
follow a professional development training targeting aspects of teacher-student relationships or choose to systematically explore possible solutions using methods such as lesson study (e.g., Cheung & Wong, 2013).

Note that successful implementation requires that teachers understand what kind of teaching behaviors are related to specific items. This goes beyond training of observers and also requires training of teachers. Currently, we unfold a project based on colleague visitation. In this project all teachers are trained and function as observers as well as participants. During observation training much time is spend on what specific teaching behaviors are indicators of an item. We developed a specific instrument for training purposes in which items are accompanied by various indicators (see Appendix).

References


Appendix: Teacher observation instrument used when training observers.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Indicator: The teacher</th>
<th>Results</th>
<th>Examples of good practice: The teacher</th>
</tr>
</thead>
</table>
| Safe and stimulating learning climate | 1 shows respect for learners in his/her behavior and language | 1 2 3 4 | … lets learners finish their sentences  
… listens to what learners have to say  
… does not make role-stereotyping remarks  
… addresses learners in a positive manner  
… uses and stimulates humor  
… accepts the fact that learners make mistakes  
… shows compassion and empathy for all learners present  |
|                                 | 2 maintains a relaxed atmosphere | 1 2 3 4 | … gives positive feedback on questions and remarks from learners  
… compliments learners on their work  
… acknowledges the contributions that learners make  
… stimulates learners to listen to each other  
… intervenes when learners make fun of someone  
… keeps (cultural) differences and idiosyncrasies in mind  
… stimulates solidarity between learners  
… encourages learners to experience activities as group events  |
|                                 | 3 promotes learners’ self-confidence | 1 2 3 4 | Learners enter and settle in an orderly manner  
… intervenes in a timely way and appropriately in case of disorder  
… safeguards the agreed rules and codes of conduct  
… keeps all learners involved in activities until the end of the lesson  
… makes sure that learners know what to do if they need help with their work and explains clearly when they can ask for help  
… makes sure learners know what to do when they have finished their work  
… checks whether learners have understood what they have to do  |
| Efficient organization          | 4 fosters mutual respect | 1 2 3 4 | … provides feedback on learners’ social functioning while carrying out a task  
… explains clearly which materials can be used  
The materials for the lesson are ready for use  
Materials are geared at the right level and developmental stage of the learners  
… starts the lesson on time  
… does not waste time at the beginning, during, or at the end of the lesson  
… prevents any unnecessary breaks from occurring  
… does not keep learners waiting  
… activates prior knowledge of learners  
… gives staged instructions  
… poses questions learners can understand  |
<p>|                                 | 5 ensures the lesson proceeds in an orderly manner | 1 2 3 4 |  |
|                                 | 6 monitors to ensure learners carry out activities in an appropriate manner | 1 2 3 4 |  |
|                                 | 7 provides effective classroom management | 1 2 3 4 |  |
|                                 | 8 uses the time for learning efficiently | 1 2 3 4 |  |
| Clear and structured instructions | 9 presents and explains the subject material in a clear manner | 1 2 3 4 | (Continued on next page)  |</p>
<table>
<thead>
<tr>
<th>Domain</th>
<th>Indicator: The teacher</th>
<th>Results</th>
<th>Examples of good practice: The teacher</th>
</tr>
</thead>
</table>
| 10     | gives feedback to learners | 1 2 3 4 | ... summarizes the subject material from time to time  
... makes clear whether an answer is right or wrong  
... makes clear why an answer is right or wrong  
... gives feedback on the way in which learners have arrived at their answer  
... creates learners assignments that stimulate active participation  
... asks questions that stimulate learners to reflect  
... makes sure that learners listen and/or continue working  
... allows for “thinking time” after asking a question  
... invites learners to participate who do not volunteer to do so |
| 11     | engages all learners in the lesson | 1 2 3 4 | ... ask questions that stimulate learners to reflect  
... checks regularly whether learners understand what the lesson is about  
... praises learners who do their best  
... makes clear that all learners should do their best  
... expresses positive expectations about what learners are going to achieve  |
| 12     | during the presentation stage, checks whether learners have understood the subject material | 1 2 3 4 | ... checks whether learners have understood the subject material  
... asks questions that stimulate learners to reflect  
... checks regularly whether learners understand what the lesson is about  
... praises learners who do their best  |
| 13     | encourages learners to do their best | 1 2 3 4 | ... encourages learners to do their best  
... praises learners who do their best  
... makes clear that all learners should do their best  
... expresses positive expectations about what learners are going to achieve  |
| 14     | teaches in a well-structured manner | 1 2 3 4 | ... teaches in a well-structured manner  
The lesson is built up in terms of clear stages and transitions between stages  
The lesson builds up logically, going from the simple to the complex  
Activities and assignments are connected to the materials presented during the presentation stage  
The lesson offers a good variety of presentation, instruction, controlled practice, free practice, and so forth.  |
| 15     | gives a clear explanation of how to use didactic aids and of how to carry out assignments | 1 2 3 4 | ... gives a clear explanation of how to use didactic aids and of how to carry out assignments  
... makes sure that all learners know what to do  
... explains how lesson aims and assignments relate to each other  
... explains clearly which materials and sources can be used  |
| Intensive and activating teaching [in short referred to as “activating teaching methods”] | offers activities and work forms that stimulate learners to take an active approach | 1 2 3 4 | ... offers activities and work forms that stimulate learners to take an active approach  
... uses diverse forms of conversation and discussion  
... offers controlled (pre-)practice  
... lets learners work in groups  
... uses Information and Communication Technology (ICT, e.g., digiboard, beamer)  
... employs a variety of instruction strategies  
... varies assignments  
... varies lesson materials  
... uses materials and examples from daily life  
... asks a range of questions  |

(Continued on next page)
<table>
<thead>
<tr>
<th>Domain</th>
<th>Indicator: The teacher</th>
<th>Results</th>
<th>Examples of good practice: The teacher</th>
</tr>
</thead>
<tbody>
<tr>
<td>17 stimulates the building of self-confidence in weaker learners</td>
<td>1 2 3 4</td>
<td></td>
<td>… gives positive feedback on questions from weaker learners</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>… displays positive expectations about what weaker learners have to achieve</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>… compliments weaker learners on their work</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>… acknowledges the contributions made by weaker learners</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>… shows learners the path they can take toward a solution</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>… teaches strategies for problem solving and referencing</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>… teaches learners how to consult sources and reference works</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>… offers learners checklists for problem solving</td>
</tr>
<tr>
<td>18 stimulates learners to think about solutions</td>
<td>1 2 3 4</td>
<td></td>
<td>… waits long enough to give all learners the chance to answer a question</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>… encourages learners to ask each other questions and explain things to each other</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>… asks learners to explain the different steps of their strategy</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>… checks regularly whether instructions have been understood</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>… asks questions that stimulate reflection and learner feedback</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>… checks regularly whether learners understand what the lesson is about</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>… provides the opportunity for learners to think aloud about solutions</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>… asks learners to verbalize solutions</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>… promotes interaction between learners</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>… promotes interaction between teacher and learners</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>… informs learners at the start of the lesson about the lesson aims</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>… clarifies the aims of assignments and their learning purpose</td>
</tr>
<tr>
<td>19 asks questions that stimulate learners to reflect</td>
<td>1 2 3 4</td>
<td></td>
<td>… evaluates whether the lesson aims have been reached</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>… evaluates learners’ performance</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>… gives weaker learners extra study time</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>… gives weaker learners extra instruction time</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>… gives weaker learners extra exercises/practices</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>… gives weaker learners “pre- or postinstruction”</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>… puts learners who need little instructions (already) to work</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>… gives additional instructions to small groups or individual learners</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>… does not simply focus on the average learner</td>
</tr>
</tbody>
</table>

Adjusts instructions and learner processing to inter-learner differences [in short referred to as “differentiation”]
<table>
<thead>
<tr>
<th>Domain</th>
<th>Indicator: The teacher</th>
<th>Results</th>
<th>Examples of good practice: The teacher</th>
</tr>
</thead>
</table>
| 26     | adjusts the processing of subject matter to relevant interlearner differences | 1 2 3 4 | … distinguishes between learners in terms of the length and size of assignments  
… allows for flexibility in the time learners get to complete assignments  
… lets some learners use additional aids and means |
| Teaches learning strategies | teaches learners how to simplify complex problems | 1 2 3 4 | … teaches learners how to simplify complex problems  
… teaches learners how to break down complex problems into simpler ones  
… teaches learners to order complex problems |
| 28     | stimulates the use of control activities | 1 2 3 4 | … pays attention to prediction strategies for reading  
… lets learners relate solutions to the context of a problem  
… stimulates the application of alternative strategies |
| 29     | teaches learners to check solutions | 1 2 3 4 | … teaches learners how to estimate outcomes  
… teaches learners how to predict outcomes  
… teaches learners how to relate outcomes to the practical context |
| 30     | stimulates the application of what has been learned | 1 2 3 4 | … stimulates the conscious application of what has been learned in other (different) learning contexts  
… explains to learners how solutions can be applied in different situations  
… relates problems to previously solved problems |
| 31     | encourages learners to think critically | 1 2 3 4 | … asks learners to provide explanations for occurrences  
… asks learners for their opinion  
… asks learners to reflect on solutions or answers given  
… asks learners to provide examples of their own |
| 32     | asks learners to reflect on approach strategies | 1 2 3 4 | … asks learners to explain the different steps of the strategy applied  
… gives an explicit explanation of possible (problem-solving) strategies  
… asks learners to expand on the pros and cons of different strategies |