Teacher evaluation through observation

van der Lans, Rikkert

Chapter 8

Discussion

8.1 Discussion

This dissertation concludes with a discussion of its overall limitations and implications (for discussions of limitations specific to each study, see the corresponding chapter). The section is divided into four broad subsections. Section 8.2 considers the more general methodological challenges encountered during this research. Section 8.3 pertains to implications for theory and discusses alternative interpretations. Section 8.4 identifies the implications for practice and discusses how schools may implement evaluation procedures. Finally, Section 8.5 discusses some directions for further research on this topic.

8.2 Methodological challenges

8.2.1 The issue of one-dimensionality

Virtually all theories presume that classroom teaching is multidimensional (e.g., Creemers & Kyriakides, 2006; Hill et al., 2012; Pianta & Hamre, 2009). Thus, it seems illogical to propose a one-dimensional cumulative order in teaching practices. Discussions concerning the dimensionality of measurement are complicated by the disagreement among statisticians about how to empirically assess one-dimensional item order. The regularly applied factor analytic approaches (for an excellent overview, see Timmerman et al., 2016) have considerable limitations, specifically if the underlying hypothesis presumes a cumulative order in item responses. Cumulative order implies that some items have few positive responses (they are complex), while other items have many positive responses (they are easy). Complex items are only responded to positively if the easy items have been responded to positively, which results in a correlation matrix with decreasing inter-item correlations (in which items of similar complexity are highly correlated, while items further apart have low correlations). In the ideal situation, the correlation between the easiest and most complex items approximates $r = 0.00$. This correlation matrix is inconsistent with the one-factor model's expected correlation matrix (Browne, 1992; Guttman, 1954; Jöreskog, 1970; Van Schuur, 2011).

Although some of the studies included in this dissertation apply factor analysis, during this research we have come to acknowledge its limitations. Our current point of view is against using factor analytic techniques based on (linear) principal components or eigenvalues to test for one-dimensional *cumulative* item ordering. Alternative coefficients may be found in the literature about Mokken scaling, such as Mokken's adaptation of Loevinger's H-coefficient (Van der Ark, 2007). Andersen's likelihood ratio (LR) test can

be applied to evaluate one-dimensionality (e.g., Chapter 4). Also, Guttman's (1954) alternative approach to factor analysis—known as the simplex factor model—seems worthy of further exploration (see Browne, 1992; Chapter 6 herein). Fox (2010) proposes to assess item local independence and claim one-dimensionality if local independence holds. Ponocny's (2001) T_1 and T_{1m} statistics or Chen and Thissen's (1997) LD χ^2 statistic could then be considered.

As a final point, although discussions about assessments of one-dimensionality are technical, they are paramount for further development of theory on teaching and teacher professional development, as well as the evaluation and measurement of teaching. Currently, researchers claim that teaching must be multidimensional because it is so complex and interactive and takes place in a dynamic environment (e.g., Creemers & Kyriakides, 2006; Pianta & Hamre, 2009). From their perspective, a one-dimensional measure of teaching skill is an oversimplification of the teaching practice itself. The studies included in this dissertation are not meant to deny the complexity of teaching; however, we argue that the complexity of teaching can be studied and visualized in two different ways. Factor analysis can be used to explore and cluster items describing teaching practices of similar complexity. For example, in the initial development of the ICALT (see Chapter 1), Van de Grift, Van de Wal, & Torenbeek (2011) used factor analysis to verify the hypothesized six domains. This approach is valuable if one aims to explore groupings of teaching practices of similar complexity in order to evaluate teachers' performance on each grouping. However, confirming that items describing various teaching practices can be ordered cumulatively in terms of complexity requires statistical models other than factor analysis. In specific, models which can confirm that items increase in difficulty and that complex items require skill in less complex items (cumulative increase). From this perspective, measurement can be multi-dimensional and one-dimensional at the same time, depending on how one wants to use the outcomes.

8.2.2 Multilevel analysis of Rasch model assumptions

While the Rasch and IRT models are widely accepted, their use has long been limited to specific kinds of data. Only recently have researchers attempted to broaden their applicability to include multilevel and multivariate data (e.g., Doran et al., 2007; Fox, 2010; Von Davier & Carstensen, 2007). This dissertation uses some of these new applications, and specifically multilevel Rasch model analysis, which is an extension of the regular

Rasch model. The regular Rasch model estimates two parameters: (1) item complexity (usually referred to as item difficulty) and (2) teaching skill (usually referred to as a person's ability). However, two parameters are too few, if, for example, students in one class rate a teacher. In this situation, many observers rate one teacher, and a more appropriate specification of the model would be to separate three parameters: item complexity, teaching skill, and observer bias. This extended specification would prevent information about observer bias and that concerning teaching skill are modeled by the same parameter. Such an extension would be a multilevel specification of the Rasch model in which the previous parameter teaching skill is subdivided into two parameters. However, while multilevel extensions of the Rasch model are available and accessible, a well-developed understanding does not yet exist about how to assess its model fit (Fox, 2010). De Boeck et al. (2011) propose some tests to evaluate model assumptions, but little information is available about the feasibility of these multilevel assumption tests. Taken together, the field is not sufficiently developed to apply these fit statistics. Therefore, the studies implement another approach that circumvent some of the problems.

The rationale underlying the approach is as follows: Applying multilevel models is appropriate generally for two reasons. First, in exceptional situations, they may give more accurate estimates of parameters involved (to prevent ecological fallacies; Hox, 2010). Second, they can provide more accurate estimates of the standard errors of the parameters. Standard errors are an indicator of unreliability and model fit. While not reported in Chapters 2, 3, and 4, the studies in this dissertation indicate little difference in the estimation of item parameters between multilevel Rasch models and the regular Rasch model. This indicates that multilevel applications are not required to prevent ecological fallacies (at least not in these studies). However, the standard errors of the item parameters are larger if a multilevel Rasch model analysis is performed. Thus, the regular Rasch model seems to underestimate the standard errors, and, as a result of this, application of the regular Rasch model fit statistics could lead to overly strict testing of model fit. Given the unavailability of multilevel fit statistics, applying regular Rasch model fit statistics seemed acceptable, provided they do not substantially rely on variance distributions or standard errors. For this reason, the studies deliberately do, for example, not use the popular infit and outfit Rasch model fit statistics (Bond & Fox, 2007) based on mean squares but instead apply the older Andersen (1973) log LR-test, which is based on the difference in item parameters between two subgroups.

8.3 Theoretical implications and considerations concerning interpretation

8.3.1 The evaluation of more complex teaching practices

The findings in Chapters 2, 3, and 4 lead us to conclude that students and observers agree on the complexity of similar teaching practices. A comparison of the results from Chapters 2 and 3 shows that teaching practices are ordered similarly, which implies that the cumulative order is corroborated by students and observers. The results in Chapter 4 go a step further by providing evidence that both students and observers fit the same one-dimensional order and that they agree on the complexity of similar teaching practices. However, the evidence supporting the agreement between students and observers is stronger for the evaluation of more basic teaching practices, including a safe learning climate, efficient classroom management, quality of instruction, and activating teaching methods domains. The evidence remains mixed with regard to the most complex teaching practices (i.e., teaching learning strategies and differentiation domains).

Thus, the results provide reasons to debate what aspects should be considered most complex in learning to teach. In Chapter 2, the classroom observers assigned differentiation as the most complex teaching competency and teaching learning strategies as substantially less complex. In Chapter 3, however, the sample of student ratings suggests that teaching learning strategies and differentiation are of similar complexity, and Chapter 4's sample reconfirms these differences between observers and students. On the basis of the studies included in this dissertation, it is tempting to explain the mixed results as due to a discrepancy in interpretation between students and observers. However, the mixed findings might not be completely dependent on the chosen type of evaluation method. For example, research in primary education has reported, on the basis of classroom observations, that teaching learning strategies and differentiation are equally complex teaching practices (Van de Grift, Van der Wal, & Torenbeek, 2011).

One explanation of the mixed findings might be found in Scriven (1981, 2007), who mentions that classroom observers' scorings are affected by common standards and norms about teaching. If this logic is valid, then various educational policy agents' calls to improve competency in the differentiation of secondary education teachers (e.g., Inspectie van het Onderwijs, 2014; CPS, 2012) might have biased classroom observers to overrate the complexity of differentiation. Though admittedly highly speculative, the call also might have legitimatized low scores on differentiation. In particular teachers feeling insecure about scoring colleagues' classroom practices or teachers confusing observation with

judgment might have searched for such legitimatizations. For them, scoring all behaviors as sufficient might have felt incorrect, but to avoid conflicts with colleagues scoring low on differentiation might have been a safe bet, according to a sense that "It is not so bad to be bad in differentiation, because many are." The only available evidence pointing in this direction is the larger number of violations of local independence found among the classroom observation items associated with the differentiation and teaching learning strategies domains. These violations are not present among the student questionnaire items associated with these domains and thus seem unrelated to the evaluation of the domains in general. The violations suggest that observers scored the items associated with differentiation as more similar than expected by the model, such that, currently, items included in the differentiation domain function too much as one item. If one item is scored unobserved, then the other items are scored this way as well. This result fits with the above speculation.

Another explanation might stem from the item content of the "My Teacher" student questionnaire. It is debatable whether items such as "connects to what I am capable of" provide a similar operationalization of the differentiation domain, compared with the classroom observation items "adapts processing of subject matter to student differences" or "adapts instruction to relevant student differences." The difference is that questionnaire items are less specific about the instructional situation. They do not specify whether the teacher connects to the student capabilities by explaining the same assignment or material at different levels of complexity or pace (adaptation of processing) or by giving the student different assignments or materials (adaptation of instruction). The questionnaire item "connects to what I am capable of" even may refer to both situations. The classroom observation instrument is more specific about such instructional differences. However, the larger number of positive residual correlations between ICALT items describing differentiation practices suggests that observers do not distinguish among them very much. This finding is inconsistent with the explanation.

Yet another explanation is that classroom observation of teaching practices included in the final two domains depends more on situational and contextual circumstances than items in the other four domains. In some lessons, teaching practices associated with the differentiation and teaching learning strategies domains are not performed, due to pedagogical choices made in advance about the design of the lesson, based on the teacher's specific educational goals (Doyle, 2006). The question now is whether this is an

appropriate choice. Should teachers always search for ways how to differentiate and teach students learning strategies or is it legitimate to sometimes chose otherwise? Related to this argument, as Kennedy (2010) points out, sometimes practices cannot be performed because of situational circumstances beyond the teacher's control. Again, violations of local independence could be expected on the basis of this explanation. It can be argued that, the explanation also is consistent with the residual correlations between items corresponding to classroom management and the items describing more complex teaching practices reported in Chapter 2. Together with the results in Chapter 6, it might be speculated that teachers incidentally (need to) choose lesson designs and educational goals that result in different classroom management procedures, which in turn obstruct the use of the most complex teaching practices. However, the student questionnaire data does not show the same patterns, and the same evidence can also be claimed to support other explanations, in particular by observation bias as mentioned previously.

Finally, students might provide invalid evaluations of teaching learning strategies. The results of Chapter 4 and, to a lesser extent, Chapter 3, in which most student questionnaire items associated with teaching learning strategies do not fit the model assumptions, provide some support for this explanation. In Chapter 4, all but one student questionnaire item measuring teaching learning strategies misfit the model assumptions, which provides reason to doubt whether students can validly evaluate teachers' use of learning strategies.

8.3.2 An alternative interpretation of the cumulative order

In this dissertation, the cumulative ordering is interpreted as reflecting teachers' personal development in teaching. However, an alternative interpretation suggests that the cumulative ordering reflects the development that the teacher and class experience across the school year. Some researchers have proposed that a safe learning climate and efficient classroom management must be established with every class at the start of the school year (e.g., Mainhard, Brekelmans, De Brok, & Wubbels, 2011) and before more complex teaching methods can be applied in that class. In this alternative interpretation, the cumulative order reflects how teacher and class learn to work together, and the teacher's developmental stage is expected to increase during the school year as a consequence of this learning. If true, the interpretation applied herein that the cumulative ordering reflects teachers' stage in their personal development is not relevant. The evidence presented in this

dissertation does not exclude this alternative explanation. One longitudinal study by Helms-Lorenz, Van de Grift, and Maulana, (2016) provides some evidence indicating that an interpretation in terms of personal development is not invalid, but this does not exclude validity of the alternative interpretation. Maybe the cumulative order reflects both. This warrants further examination in future studies.

8.3.3 An alternative explanation for the consistency of the ordering

The studies in this dissertation are field studies. They examined teachers' professional development in schools and did not attempt to experimentally manipulate any aspects of teachers' professional development, which somewhat hampers any firm conclusions that all teachers must develop according to these stages. We acknowledge the possibility that the field has some common latent norms about how teachers should learn the profession. If this is the case and teachers are educated using similar didactics, it might explain the consistency in the development of teaching skill and the lack of different development paths (Chapter 6).

Participating teachers noted considerable overlap between the teaching practices mentioned in the ICALT and "My Teacher" instruments and standard works used by many Dutch teacher education institutes, such as Ebbens and Ettekoven (2005) and Teitler (2013). Thus, there are grounds to argue that teachers already in Teacher Education start learning the profession in a manner similar to the cumulative order established in this dissertation. While, this alternative explanation strengthens the validity of the presented results, it also presents reason for caution. The lack of individual differences reported in Chapter 6 might be explained by the similarities in teachers' background education; that is, virtually all teachers have spent the most time learning how to secure a safe learning climate, efficient classroom management, and how to provide understandable classroom instructions. Similarly, virtually all teachers have spent considerably less time learning how to teach students learning strategies and in differentiation strategies. Thus, the results do not completely exclude the possibility that, for example, teachers could achieve skill in differentiation before they acquire skill in classroom instructions if they would have received more time to train and learn teaching practices associated with that specific domain. While this explanation is theoretically interesting, it has low practical utility for schools and teachers. Even if an experimental manipulation can show that teachers could develop their profession differently, the evidence is clear that in practice they do not do so.

8.3.4 Expected impact on student learning

The developed instruments are grounded in literature about teaching effectiveness. An important question is what can be expected when teachers succeed in improving their skill. Are teachers who successfully implement more complex teaching practices also more effective? Van de Grift and Lam's (1998) empirical study addresses the predictive validity of the ICALT, showing a significant positive effect on student achievement in primary education. Furthermore, we note that ICALT and "My Teacher" show much overlap with other instruments currently in use, including the Classroom Assessment Scoring System (CLASS) and Framework for Teaching (FFT) (e.g., Maulana et al., 2015). Other studies show that these instruments are predictive of student achievement gains (Kane et al., 2012).

The research performed for this dissertation was also meant to further support the assertion that higher scores on the ICALT observation and "My Teacher" questionnaire are related to student learning and school success. To this end, we gathered teacher-assigned grades of all participating teachers (rather than normative achievement tests, because Dutch secondary education uses normative achievement tests only for final exams [since very recently some schools also yearly evaluate progress in reading and math using normative tests]). However, analyses revealed that teacher-assigned grades are too unreliable to identify differences in effectiveness between teachers (see the Appendix A). Therefore, we made no further attempt to connect teacher-assigned grades to ICALT and "My Teacher" evaluation outcomes.

8.4 Practical implications and considerations for use

An important advantage of the ICALT and "My Teacher" instruments is their capability to provide teachers with diagnostic information about current performance and the most promising directions for further teacher training. The cumulative order established and validated in Chapters 2, 3, and 4 offers great potential to contribute to the provision of feedback. The main advantage of a cumulative item order that reflects complexity levels in teaching is that it can be used to scaffold feedback to the appropriate level of skill. Specifically, areas whose complexity are near the teacher's skill are most relevant for further training and professionalization, whereas both more and less complex areas are less relevant. Therefore, using this cumulative ordering can point to the most plausible ways individual teachers can improve their teaching.

To use such diagnoses effectively, however, it is necessary to recognize that any diagnosis is rather uninformative by itself. That is, providing a teacher with feedback about improving use of teaching methods is in itself of little use. Using diagnoses effectively requires that teachers, coaches, schools, and teacher educators build a knowledge base about how best to act on a specific diagnosis. For example, the item "ensures mutual respect" functions as an umbrella under which a range of behaviors can be specified. Advice might include reading theory about teacher–student relationships (e.g., Wubbels & Brekelmans, 2005; Pianta & Hamre, 2009) or discussing possible strategies with colleagues. Alternatively, the teacher could choose to follow a professional development training targeting aspects of teacher–student relationships or systematically explore various interventions using methods such as lesson study (e.g., Ming & Wong, 2013).

Another important condition for successful implementation requires that teachers understand which behaviors are related to specific items describing a specific teaching practice. In the typical research setting, observers are only trained in how to interpret items. Ensuring that feedback is understandable to teachers goes beyond training observers to include training teachers.

8.4.1 How to organize teacher feedback in schools

For this dissertation, data were gathered within a specific evaluation procedure: Schools grouped teachers into teams of four, and teachers within a team observed one lesson of each of their team members. Thus, the data set consisted of three classroom observations for each teacher, the necessary number of lesson visits according to extant research required to obtain modest reliability (Hill et al., 2012; Kane et al., 2012). In addition, the study design required that the team of teachers teach to the same class to ensure that collegial observers would be familiar with student behavior and could notice and learn from any differences. As such, teacher evaluation might already stimulate teacher learning and professional development (Van Veen, Zwart, Meirink, & Verloop, 2010).

In addition, from an organizational point of view, it was necessary to use anonymous scores from individual observers: If a teacher received feedback or evaluative decisions after one observer visited the lesson, classroom observation is not anonymous. This makes the observer vulnerable to criticism (French-Lazovik, 1981; Peterson, 2000; Scriven, 1981). In this one-lesson visit procedure, observers might choose to avoid conflicts by giving overly lenient scores. Centra (1975) and Weisberg et al. (2009) provide some

evidence in support of this. Because the evaluation procedure applied herein provides feedback based on multiple observers, it is impossible to blame any specific observer, which could bolster observers' confidence in accurately scoring the teaching practices observed.

Finally, from the scientific point of view, the requirement that the team of teachers teach to the same class is imposed because teaching effectiveness is known to vary between classes, and this variation is not entirely due to differences in teaching (e.g., Goldhaber & Hanssen, 2013). Having information from multiple teachers within the same class allows evaluators to take such variation into account. However, the resulting evaluation procedure is considerably more complex than the standard procedures using a single class or a single classroom observation.

8.4.2 Once is not enough: The need for multiple lesson visits

The results in Chapter 5 corroborate previous findings regarding reliability (Hill et al., 2012; Kane et al., 2012). They suggest that a single classroom observation does not provide enough information about the teacher's general skill. Therefore, schools willing to invest in teacher evaluation should implement evaluation procedures that use different peers visiting lessons, with a minimum of three visits, and they should not provide feedback on the basis of the single observations. Providing feedback on the basis of single classroom observations presents the risk of inaccurate feedback that will not improve teaching and result in wasted resources for teacher training on inadequate professionalization trajectories, coaching, courses, or schooling. In addition, it might demotivate the teacher.

8.4.3 How to provide direct feedback

Using the procedure as described in the preceding section, would deny individual observers the opportunity to give direct feedback, which considerably constrains teachers' learning opportunities. Strictly speaking, the peers leave without providing feedback, and teachers receive it only after three peers have visited, they do not receive direct feedback.

Some nuance is required here. Reliability involves the degree to which scores can be generalized to other situations, and our findings indicate that evaluation outcomes based on one-time classroom observations do not provide information about the teacher's *general* teaching skill. Nevertheless, they offer reliable insights about the specific lesson observed. Colleagues can provide effective direct feedback if they (1) do not rely on or mention

specific item scores (to secure anonymity) and (2) give feedback about the specific lesson and avoid claims about the teachers' general teaching skill. Scriven (1981) mentions that one-time lesson visits can only be used to give feedback about how the teacher reacted in specific situations during the lesson, such as a misbehaving student or a student question. We disagree with Scriven (1981) that such situations do not occur often or are unimportant. Teachers view their profession as highly idiosyncratic and often are most concerned about (and look for reassurance on) how they reacted to a "difficult" student or question.

8.4.4 How to organize evaluative decisions in schools

Another important consideration is how to organize evaluative decisions in schools. Some have proposed that evaluative decisions should be organized separately from feedback (Peterson, 2000; Popham, 1987; Scriven, 1968). As Popham (1987) and Peterson (2000) caution, if data obtained for the purpose of providing teachers with feedback are also used for evaluative decisions, teachers likely would be reluctant to admit to any specific situations in which they feel incompetent, the very situations in which they are most in need of feedback. However, organizing two separate evaluation procedures also is inefficient, because it requires different personnel and protocols for each procedure, and long-term implementation is unlikely, given schools' limited resources and time. Therefore, this dissertation proposes to extend the current procedure for feedback if the aim is to make evaluative decisions. The extended procedure includes more classroom observations and combines them with information obtained with other measures of teaching skill. However, the observations used for feedback can be included in this number. This approach would avoid two separate evaluation procedures; yet it might provide sufficient confidence and protection such that teachers feel safe to share their difficulties.

The results in Chapter 5 indicate low gains in terms of reliability if schools gather numbers of lesson visits beyond 10 (less than .01 increase in reliability). If schools gather 10 lesson visits, reliability is estimated as .83. Hence, the required level of .90 reliability is beyond reach if schools only collect classroom observations of teaching. Thus, schools need to gather additional information to further lower the chances of wrongly offering tenure to or dismissing a teacher. We propose gathering yearly four observations in combination with one student questionnaire, to use this data to set teacher's learning goals as well as to provide feedback and to repeat this cycle for three years before schools use the combined information of 12 observations and 3 student questionnaires in a performance

evaluation. This guarantees reliability levels which are certainly above .85 and likely above .90. Using such strategies teachers receive reliable feedback from both colleagues and students for three years in a row and are evaluated the fourth year to receive tenure or payment adjustments.

The results stress the need for carefulness if making summative evaluative decisions. Evaluative decisions cannot be grounded on classroom observations only (Kane et al., 2012; Peterson, 2000). It seems that, at a minimum, classroom observations should be complemented with student questionnaire results, not only to further increase reliability to an acceptable level, but also to add validity and trust in the evaluation outcomes. Schools could also consider adding other evaluation methodologies. For example, Peterson (2000) proposes giving teachers themselves a voice in the evaluation methodology to increase their confidence in the evaluative decisions. Alternatively, schools might consider implementing some mandatory and some optional evaluation methods.

8.5 Further research: Where to go from here?

8.5.1 Expected applicability of the Rasch model to other instruments

As Chapter 1 explains, the validated theory predicts a cumulative ordering of development in teaching skill. We chose the Rasch model over other statistical models because it is particularly powerful in identifying cumulative response patterns. If the theory is valid, the cumulative item order should also apply to other instruments if they include items describing various effective teaching practices. Therefore, it is relevant to apply the Rasch model to data obtained with other observation instruments and questionnaires currently employed in schools (e.g., Overdiep, 2016). Such an analysis might further validate the presented theory of cumulative teacher development, and it also provides opportunities to validate other currently employed instruments to evaluate teaching (e.g., Overdiep, 2016).

8.5.2 Resolving the Dutch challenge: How to envision national teacher evaluation

Chapter 1 explains that teacher evaluation in the Dutch context needs to be organized differently from teacher evaluation in many other Western countries, because the Dutch school system is characterized by high school autonomy and high teacher autonomy, which precludes any hierarchical implementation of national evaluation procedures. The lack of a nationwide evaluation procedure does not result in a lower frequency of evaluation though. Dutch Inspectorate (Inspectie van het Onderwijs, 2016) and OECD (2016) reports suggest

that many schools evaluate teachers yearly, which makes the Netherlands among the top evaluators in terms of frequency (Isoré, 2009). More problematic is the large amount of untested observation and questionnaire instruments that are applied (Overdiep, 2016). The applied evaluation procedures frequently involve only one classroom observation or one questionnaire. Overdiep (2016) creates some awareness that current evaluative decisions and feedback frequently lack empirical support and that schools should start using instruments that, if implemented properly, *can* result in valid and reliable decisions and feedback. However, given the autonomy in Dutch schools, even if schools chose to apply only scientifically validated instruments, they still would likely choose different instruments. Thus, the diversity in instruments used by different schools will continue to complicate comparisons across schools and teachers.

Additional research might examine whether these different instruments can be linked using IRT-linking designs (e.g., Vale, 1986; Weeks, 2010). In a linking design, all schools can use their own instrument, but all instruments are linked together into one larger instrument. The advantage is that schools have the freedom to formulate their own items for evaluation but still gain comparable feedback and evaluations. To link different observation instruments and questionnaires effectively, it is advisable, though not strictly necessary, that all instruments have a few items in common. These anchor items can be used to estimate the position of the other items. Furthermore, items should fit the Rasch model assumptions to ensure valid and reliable evaluation.

8.6 To conclude

In conclusion, this dissertation provides evidence of the validity of the ICALT classroom observation instrument and "My Teacher" questionnaire to support feedback and evaluative decisions about teaching skill. It also provides evidence supporting a theory of stagewise development in effective teaching practices and shows that this ordering can be studied using different evaluation methods. In addition, it discusses the relevance of the evaluation procedure to guarantee sufficient evaluation reliability. In the future, researchers will need to consider how to build on these results to ensure valid and reliable teacher evaluation in all schools.