Chapter 7

Conclusions

7.1 Conclusion

This dissertation assesses various aspects of the validity of a theory predicting cumulative development in teaching practices (Chapters 2, 3, and 4). To do so, the studies use two instruments specifying a wide range of teaching practices: the ICALT classroom observation instrument and the "My Teacher" questionnaire. Validating this theory is important because of its wide application in various institutes and projects across the Netherlands to provide teachers feedback, as well as to evaluate professional skill. In addition, Chapters 5 and 6 investigate the reliability of feedback and evaluative decisions. Establishing this reliability for individual teachers is critical; if reliability is low, there is a greater likelihood that they receive inaccurate feedback or that school principals make inaccurate evaluative decisions. Unreliable evaluation does not improve teachers' skill and can even harm them. The following subsections present an overview of the main conclusions and findings of each chapter, before summarizing the general conclusions.

7.2 Research questions, by chapter

Chapter 2. Can classroom observations of effective teaching practices be ordered cumulatively? And; what does this ordering learn us about the development of effective teaching?

Chapter 2's study results confirm that 31 of the original 32 effective teaching practices exhibit a cumulative ordering and that the ordering strongly parallels Fuller's (1969) stages. In addition, the study results further corroborate findings from Van de Grift et al.'s (2014) study, which assesses the validity of the same observation instrument for evaluating beginning teachers (less than 3 years of experience). We therefore suggest that this ordering describes a stagewise development of effective teaching practices. This development begins by developing practices to achieve a safe learning climate, proceeds with development of teaching practices directed at an efficient classroom management and quality in instruction, and finally ends with developing practices in domains related to activating teaching methods, teaching learning strategies, and differentiating and adapting lesson content to meet student needs.

Chapter 3. Can student questionnaire ratings of effective teaching practices be ordered cumulatively? How may the development of such a scale contribute to the knowledge about teacher development?

Chapter 3's results confirm that effective teaching practices can be ordered cumulatively, from basic to more complex. The results further confirm those of Maulana et al. (2015), who use the same questionnaire with a sample of beginning teachers and establish a cumulative ordering similar to that presented herein. Broadly, the cumulative ordering observed is in accordance with Fuller's (1969) theory on teacher development, which states that teachers are first concerned with the self, then with the task, and finally with their impact on student learning. Thus, the validation of a cumulative ordering provides some initial insights into the development of effective teaching practice.

Chapter 4. To what extent do observers and students agree on the cumulative ordering in teaching practice complexity?

Chapter 4's study combines observation instrument (the ICALT observation) and student questionnaire (the "My Teacher" questionnaire) data and explores whether items of both instruments measure one latent variable, namely, teaching skill. The results indicate that, in general, students and observers agree on the complexity of similar teaching practices and order them similarly, adding some additional insights to Maulana and Helms-Lorenz (2016). They confirm the moderate correlation (r = .26) between evaluation outcomes based on a single classroom observation and student ratings on a questionnaire. However, Chapter 4's study disconfirms previous speculations that this low correlation can be explained by differences between students' and observers' interpretations of items. Without doubt, student questionnaires can address questions that observers cannot readily observe (e.g., whether students understood the explanation). Similarly, classroom observation can evaluate aspects of teaching skill that students cannot reasonably evaluate (e.g., the quality of the lesson content and materials). However, our results suggest that when observers and students evaluate aspects of teaching they both can observe, their responses to items are psychometrically similar and one-dimensional.

Chapter 5. How many classroom observations by peers are required to achieve modest reliability and support formative feedback? And; How many classroom observations by peers are required to achieve high reliability and support summative decisions?

Chapter 5's results indicate that reliable formative feedback demands observations of at least four different lessons by different peers. Also, results indicate that if 10 lesson observations are gathered, the predicted reliability is .83 and further increasing the number of gathered lessons observations is predicted to hardly increase reliability further (predicted increase smaller than .01). For summative decisions it seems required to combine lesson observations with other types of information about teaching, including student questionnaires, achievement gains, or teacher self-report. The results align with previous findings that predict modest reliability when three to four different observers visit one another's lessons (e.g., Hill et al., 2012; Kane et al., 2012; Ho & Kane, 2013). The findings share some similarities with results from five other classroom observation instruments used in previous studies (Hill et al., 2012; Ho & Kane, 2013; Kane et al., 2012), including the classroom assessment scoring system (CLASS), the framework for teaching (FFT), the UTeach observation protocol (UTOP), the mathematical quality of instruction (MQI), and the protocol for language arts teaching observation (PLATO). Chapter 5's study adds to these works by showing that these reliability coefficients also can be achieved with less complex evaluation procedures and without overly restrictive training protocols. The value of at least four for modest reliability therefore is highly relevant for real-world evaluation practices and research. Also, the finding that classroom observations alone are insufficient to guarantee acceptable reliability for summative decisions supports the overall consensus that reliable and valid teacher evaluation requires a combination of various measures. The study provides preliminary insights for how to implement classroom observations using cost-effective, manageable procedures while still ensuring generally acceptable reliability.

Chapter 6. How many observed lessons show substantial deviation from the cumulative ordering? And; Do deviating lessons cluster with some particular teachers?

Chapter 6's results suggest that approximately 15% of the lesson observations show substantial deviations from the predicted cumulative order. Further exploration indicates that misfit of the cumulative order seems not to be specific to some teachers. Only three

teachers repeatedly deviated from the predicted order on two (out of three) lessons observed. Thus, observations showing misfit seem to involve incidental lessons that could have been taught by any teacher. The results corroborate Berliner's (2001) and Sternberg and Horvath's (1995) observations that some lessons are "exceptional" and do not fit the general developmental sequences. However, the results also go beyond this observation and show that such exceptional lessons are not typical to specific teachers. In so doing, the results disconfirm speculations about individual differences in the development of teaching. Given this result, it seems reasonable to apply the same stage theory to evaluate all teachers. However, incidental observations showing misfit should be identified and removed to avoid biased feedback or biased evaluative decisions.

7.3 Overview of general findings

Overall research question: *How can classroom observation instruments and student questionnaires provide teachers and schools with valid and reliable feedback and evaluative decisions?*

The evidence suggests that classroom observations and student questionnaire ratings can be used to provide teachers and schools with valid and reliable feedback and evaluative decisions. The samples studied in this dissertation provide consistent evidence of a cumulative order: The teacher begins with learning teaching practices associated with safe learning climate and ends with learning teaching practices associated with differentiation. This ordering was evident using two different evaluation methods: classroom observation and student questionnaire ratings. Furthermore, this ordering aligns with teacher professional development theory, which warrants an interpretation in terms of professional development. Schools and teachers might use this cumulative order to inform teachers about their current stage in development and provide feedback about what has already been acquired, what can be developed and learned now, and what is yet too difficult to learn.

However, the evidence of validity is limited in some ways. First, the study uses only the "My Teacher" questionnaire and the ICALT observation instruments; therefore, it remains uncertain whether the validity of the theory generalizes beyond these instruments. Second, the supporting evidence is stronger for less complex teaching practices, which most teachers develop at the beginning of their professional careers. The cumulative ordering of these less complex teaching practices is highly consistent across studies and

evaluation methods. For the most complex teaching practices though, findings on the cumulative ordering show slight variation between the evaluation methods (see Section 8.2.1).

Regarding the reliability of evaluations, the evidence indicates that feedback and evaluative decisions based on one-time lesson visits provide unreliable evaluations of teachers' teaching skills. If schools choose to use one-time lesson visits to evaluate teaching, they take the risk of providing individual teachers with inaccurate feedback and making inaccurate evaluative decisions. The evidence is based on data obtained with the ICALT observation instrument, and strictly speaking, the conclusions cannot be generalized to other observation instruments. However, given the findings' considerable consistency with previous research based on five other observation instruments (Hill et al., 2012; Kane et al., 2012), it seems acceptable to conclude that the results are typical to classroom observation methodology in general. More complicated evaluation procedures, in which multiple observers visit multiple lessons, are necessary to increase reliability to acceptable levels. The results presented herein suggest that 4 lesson observations by four different observers are required to achieve acceptable reliability for feedback and that lesson observations need to be combined with other evaluation methods to achieve a reliability level acceptable for evaluative decisions.