Teacher evaluation through observation

van der Lans, Rikkert

Teacher evaluation through observation

Application of classroom observation and student ratings to improve teaching effectiveness in classrooms

rijksuniversiteit groningen / faculty of social and behavrioal sciences / teacher education

ico
Interuniversity Center for Educational Research

Title: Teacher evaluation through observation

Print: Ipskamp printing

Cover design: Wilco Tomassen

ISBN/EAN: 978-90-367-9683-5

university of
groningen

Teacher evaluation through observation

Application of classroom observation and student ratings to improve teaching effectiveness in classrooms

PhD Thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus Prof. E. Sterken
and in accordance with
the decision by the college of Deans.

The Thesis will be defended in public on

Monday 15 May 2017 at 14.30 hours

by

Rikkert Martijn van der Lans

born on 21 december 1984
in Ermelo

Supervisors
Prof. W. J. C. M van de Grift
Prof. K. van Veen

Assessment committee
Prof. M.P.C. van de Werf
Prof. R.R. Meijer
Prof. I. F. de Wolf

Table of contents

Chapter 1

General Introduction

1.1 General introduction

Since the turn of the century, the implementation and improvement of teacher evaluation has been a global challenge for educational policy (DfEE, 2012; Inspectie van het Onderwijs, 2016; Isoré, 2009; Mourshed, Chijioke, & Barber, 2010; National Council on Teacher Quality [NCTQ], 2013; Organisation for Economic Co-operation and Development [OECD], 2016). Research consistently shows that some teachers are more effective than others (e.g., Hanushek, 2011; Hattie, 2009; Marzano, 2003). This difference denies equally intelligent students an equal opportunity to succeed in schools, which in turn affects their future career opportunities, for better or worse. Therefore, educational policy makers have turned to teacher evaluation to diagnose differences in teaching effectiveness and to provide incentives and feedback to teachers with which to improve their lessons.

The increasing attention given to teacher evaluation affects individual teachers. Their teaching performance is monitored more frequently than in the past, and the resulting evaluative decisions can have substantial consequences. This increasing importance in turn has stimulated debate about how to protect teachers against unfair evaluations (e.g., Darling-Hammond, 2013; Peterson, 2000; Winters & Cowen, 2013). Individual teachers have a great deal at stake; they typically have worked hard to earn accreditation and to succeed in classrooms. Researchers and policymakers thus have an obligation to consider the validity and reliability of their feedback and decisions carefully, because invalid, unreliable feedback will not improve teaching effectiveness and invalid unreliable evaluative decisions could deny effective teachers promotion or tenure.

The studies included in this dissertation attend to two issues central in current debates about teacher evaluation: how to acquire valid feedback for teachers, and, secondly, how to ensure that evaluations of particular teachers' teaching practices are reliable. Valid feedback requires an understanding how teaching develops. However, evaluation measures typically are dedicated to the identification of effective teaching and generally lack such an understanding. Therefore, the studies included in Chapters 2, 3, and 4 examine how effective teaching develops, and discuss how this development can be used to scaffold feedback to individual teachers. Second, valid feedback and valid evaluation in general also requires that we reliably diagnose current teaching proficiency. The studies included in Chapters 5 and 6 investigate reliability of feedback and evaluative decisions based on classroom observations, since this method is generally considered most promising (Darling-Hammond, 2013, Marzano & Toth, 2013; Strong, 2011). Together the studies contribute to existing knowledge about how schools and policy makers can organize teacher evaluation within schools such that the resulting feedback and decisions will be valid and reliable and can be expected to increase teaching effectiveness. The overarching research question addressed is as follows:

> *How can classroom observation instruments and student questionnaires provide teachers and schools with valid and reliable feedback and evaluative decisions?*

1.2 Teacher evaluation: Context and (inter)national developments

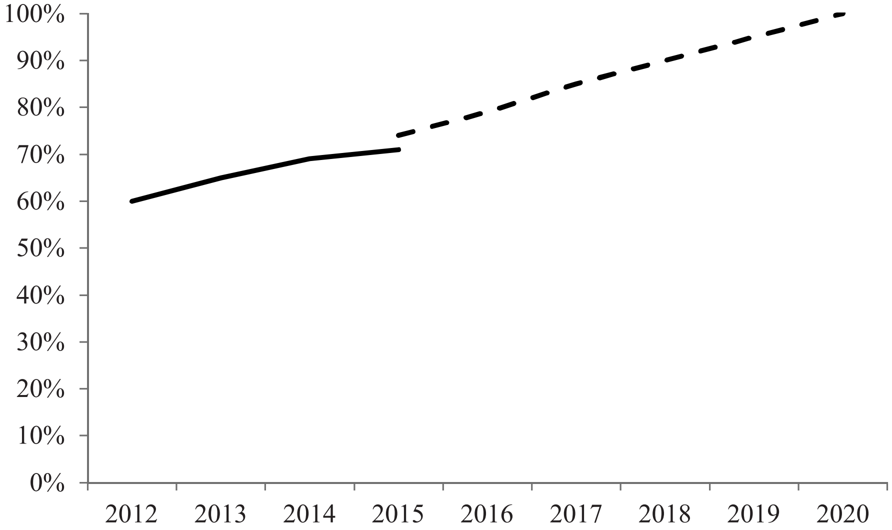
The studies addressed herein took place in the Netherlands. Organizing teacher evaluation in the Dutch context involves a relatively unique challenge compared with other countries, in that educational policies allow schools considerable autonomy in organizing teacher evaluation and grant teachers similar autonomy in their teaching (OECD, 2016; Nusche, Braun, Halász, & Santiago, 2014). The autonomy provided to the schools restricts any implementation of national hierarchical evaluation procedures such as those used in, the United States, England Germany, and France (e.g., DfEE, 2012; Hazi & Rucinsky, 2009; Isoré, 2009; U.S. Department of Education, 2009). However, in comparison to countries that provide likewise autonomy to the schools (e.g., Finland; Murillo, 2007), Dutch teachers are granted considerable autonomy. Finland, for example, has a national curriculum (e.g., Vitikka, Krokfors, & Hurmerinta, 2012), thus allowing teachers less freedom to decide how to teach. So, challenging of the Dutch context is that schools autonomously choose their evaluation methods and instruments and use them to evaluate classroom teaching of relatively autonomous teachers. Because schools are autonously choose their evaluation methods, methods and instruments vary considerably from school to school. This context makes it difficult to gather large scale data obtained with similar evaluation methods and procedures. Also, the research focus on occasion is specifically directed to how schools and teachers (instead of districts or states) may choose to organize teacher evaluation.

Another part of the context that shaped the focus of our research concerns the by the Dutch government launched policy program “de lerarenagenda” [the teacher agenda] (OECD, 2016; Nusche, et al 2013). As part of this agenda schools are required to increase their evaluation frequency. Since 2012, this policy has led to a 10% increase in the number of secondary education teachers yearly receiving a performance evaluation: from 60% to

71% (see Figure 1.1). Government ambitions are to increase this percentage to 100% in 2020.

Figure 1.1

The number of Dutch secondary education teachers receiving yearly performance evaluations. The solid line shows the increase that has been realized, the dashed line shows the ambition.



Source. Dutch ministry of Education, Culture and Science [OCW] (2016, November). https://www.delerarenagenda.nl/de-lerarenagenda/scholen-als-lerende-organisaties

The attention given to performance evaluations is observed across the globe and is, thus, not typical to the Dutch context (e.g., DfEE, 2012; National Council on Teaching Quality [NCTQ], 2013). However, the focus on performance evaluations contrasts with previous Dutch policy which typically did not have this focus (Nusche, Braun, Halász, & Santiago, 2014). And, while in general current performance evaluations in the Netherlands still do not result in decisions regarding payment, tenure, or dismissal, schools are autonomous and there is no guarantee that specific schools do not make such decision or might not start to do so in the future. Already schools are required to further differentiate in payment (commonly referred to as the “functiemix”) (Nusche et al., 2014), thereby requiring schools to make evaluative decisions regarding payment. So, it seems that the context of evaluation is changing, and it cannot be expected that instruments developed for teacher evaluation will only be used to provide formative feedback. Therefore, Chapters 5 and 6 outline two criteria for reliability: one somewhat lower criterion considered acceptable for feedback and one higher criterion deemed acceptable if evaluations are included as evidence to support high-stake decisions. Also, these studies explore some requirements to guarantee sufficient reliability.

Furthermore, the practice of peer review and feedback is also promoted and stimulated as part of the teacher agenda (e.g., OCW, 2013a; Nusche et al., 2013). Nusche, et al. identify several challenges for peer review and feedback one of which concerns the connection between evaluation and teachers' professional and career development. Chapters 2 and 4, therefore, study observations of effective teaching practices by peer-colleagues and examine stages in the development of effective teaching to better connect observations of current teaching with specific advice for professional development.

The focus of our research is further shaped by the limited availability of empirically tested evaluation instruments that can be used on a small scale within a school. This problem is not unique to the Dutch context; research articles and policy documents across the globe have noted the dearth of classroom observation instruments (e.g., Kane et al., 2012; Overdiep, 2016; Patrick & Mantzicopolous, 2016) and student questionnaires (Bill & Melinda Gates Foundation, 2012; Isoré, 2009) for which the resulting feedback and evaluative decisions have been thoroughly empirically tested or validated. In the Netherlands, this problem results in schools using various, untested observation and questionnaire instruments. Recently published information from PO-raad (representatives of the boards of all primary schools in the Netherlands) identified 33 different evaluation instruments currently applied by Dutch primary schools (Overdiep, 2016), many of which lack empirical support and some of which were developed by the schools themselves. Therefore, the studies in this dissertation focus on the validity of feedback and evaluative decisions made on the basis of two instruments: one classroom observation and one student questionnaire.

1.3 Evaluating the quality of evaluation: The conceptualization of validity and reliability

Validity and reliability are scientific concepts used to assess the appropriateness of proposed interpretations and use of scores (Kane, 2006, 2013). In the context of teacher evaluation, scores are interpreted and used in terms of feedback and evaluative decisions informative to individual teachers. The following subsections discuss the conceptual understanding of validity and reliability.

1.3.1 Validity

Across time, researchers have proposed, discussed, and disputed many different types and definitions of validity. Kane (2006, 2013) provides a comprehensive review of the development of the concept. Currently, the consensus is that validation pertains to the interpretation and use of scores obtained with an instrument or test (e.g., Cronbach, 1988; Shepherd, 1993; Bachman, 2002; Kane, 2013), which suggests a major shift from the traditional focus on validating the instrument or test itself (e.g., Cronbach & Meehl, 1955; Thorndike, 1918). This emphasis clarifies that validity involves the theory, the resulting interpretation and use, and the underlying assumptions rather than the instruments with which the theory happens to be investigated. In other words, any instrument is valid, but the interpretation given to its results and/or the way the instrument is used may be invalid.

This shift has several implications. First, evaluation instruments cannot be used for any given purpose, but only for those purposes that fit the empirically supported interpretation and use. Second, researchers must be explicit about the intended interpretation and use of their instruments. Third, if the proposed interpretation and use are valid, providing empirical support using different instruments, methods, and statistics should be possible. Fourth, whether a specific interpretation or use is valid is a matter of degree. Empirical evidence always has flaws and is generally based on small samples. Therefore, replication becomes extremely important and should involve different instruments, methods, and statistical analyses. This dissertation replicates the validity of the proposed theory and interpretation using different samples and methods (i.e., classroom observation and student questionnaire), as well as different statistical methods (i.e., item fit statistics and person fit statistics) (see Chapters 2, 3, 4, and 6).

1.3.2 Reliability

Researchers have also extensively discussed reliability, as summarized by Cronbach (2004) and Brennan (2004). Traditionally, reliability is conceptualized as the repeatability or replicability of research results (e.g., Spearman, 1904, 1910). In teacher evaluation, an exact estimation of the replicability would require an observer to do exactly the same lesson visit twice or a class of students to rate the teacher twice at the exact same moment. The similarity in feedback and/or evaluative decisions would then indicate their replicability. However, such exact repetition is impossible; a teacher cannot redo a specific lesson, such that an observer can observe exactly the same lesson twice. Moreover, exact replication has little utility (Cronbach, Gleser, Rajaratnam, & Nanda, 1972). Teachers and evaluators are usually not interested in whether an observer would evaluate the same lesson similarly; their interest is typically broader than a single lesson, which represents only a small sample of a teacher's skill. Furthermore, they are interested in the teacher's skill as observed by others, and one observer is only a very small sample of many possible others.

In this study, feedback or evaluative decisions have high reliability if they are not expected to change much if other lessons would have been visited or if other observers would have visited the lessons. Therefore, reliability is conceptualized as the generalizability of feedback and evaluative decisions, analogous to Cronbach et al.'s (1972), Shavelson and Webb's (1992), and Brennan's (2001) descriptions. In other words, reliability indicates whether repeating the evaluation procedure is likely to result in similar feedback and evaluative decisions. Chapter 5 elaborates on this concept with an example, showing that an evaluation procedure in which one observer visits one lesson has low reliability because it leads to feedback and evaluative decisions expected to change substantially if the procedure is repeated with another observer observing another lesson. The chapter also demonstrates that reliability increases when multiple observers observe multiple lessons of the same teacher.

1.4 Theory behind the instruments: Proposed interpretation

The studies herein use two different evaluation instruments: the International Comparative Analysis of Learning and Teaching (ICALT) observation form, initially developed at the Dutch Inspectorate of Education (Van de Grift, & Lam, 1998, Inspectie van het Onderwijs, 2009; Van de Grift, 2007), and the "My Teacher" questionnaire. Currently, these instruments are used in various national and international projects, including the Dutch national teacher induction project (OCW, 2013b), a regional project focusing on improving teaching at low-performing schools (Van de Grift, 2013), and the international ICALT3-project, in which instrument properties are compared globally. Furthermore, the department of Teacher Education at University of Groningen has adopted the ICALT as an instrument to coach and assess the quality of its student teachers. Note that the exact content of the

ICALT varies somewhat between institutes, in this dissertation, “ICALT” refers to the instrument as formulated by Van de Grift (2007, 2014).

The studies included in this dissertation assess the plausibility of the current routine interpretations of scores obtained with the ICALT observation instrument and the “My Teacher” questionnaire. Initially, both instruments were constructed to measure differences in teaching effectiveness (Van de Grift, & Lam, 1998; Inspectie van het Onderwijs, 2009; Van de Grift, 2007, 2014). The teaching practices included were selected on the basis of several studies, reviews, and meta-analyses showing that these practices are related to higher student achievement (gains) (e.g., Creemers & Kyriakides, 2006; Hattie, 2009; Kyriakides, 2013; Marzano, 2003; Muijs, Kyriakides, Van der Werf, Creemers, Timperley & Earl, 2014; Van de Grift, 1990, 2007, 2014). In this initial phase, developers used other criteria, such as identifying items that could be grouped into the underlying six domains, confirming that items included in a domain were internally consistent (Van de Grift & Lam, 1998; Van de Grift, 2007, 2014), and determining whether evaluation outcomes are predictive of student achievement gains, as examined by Van de Grift and Lam (1998), who show that observational outcomes are positively related to student achievement in primary education, and by Maulana, Helms-Lorenz, and Van de Grift (2015), who find positive relationships between observational scores and student engagement. This evidence all supports interpretations of differences in teaching effectiveness.

In the national induction project and the regional project focusing on low-performing schools, the instruments are interpreted as measuring teachers’ development in effective teaching practices (e.g., Helms-Lorenz, Van de Grift, & Maulana, 2016; Van de Grift, Helms-Lorenz, & Maulana, 2014; Van de Grift, Van de Wal, & Torenbeek, 2011). This interpretation connects the evaluation results with theory on teacher development, particularly Fuller’s (1969) stage theory of teacher concerns. It argues that for all teachers, development of skill in teaching can be approximately described by six cumulative stages: (1) learning how to establish a safe learning climate; (2) learning how to efficiently manage a classroom; (3) developing skills in instruction; (4) developing skills in more advanced teaching methods, including methods to activate students; (5) learning how to teach students learning strategies; and (6) developing skills in differentiation of instruction (Figure 1.2).

In this second interpretation, the scores obtained with the ICALT and “My Teacher” reflect teachers’ current stage of development in teaching skill. In addition, this interpretation predicts that these domains (or stages) fit the same one-dimensional cumulative ordering. Although the evidence provided in this dissertation can be used to support both interpretations, it focuses more on the second interpretation than the first.

Figure 1.2

Staged progression of development in teaching skill

	climate	management	instruction	activation	strategies	differentiation
Least effective teaching	✓	✗	✗	✗	✗	✗
	✓	✓	✗	✗	✗	✗
Average effective teaching	✓	✓	✓	✗	✗	✗
	✓	✓	✓	✓	✗	✗
Most effective teaching	✓	✓	✓	✓	✓	✗
	✓	✓	✓	✓	✓	✓

In Figure 1.2, the check boxes indicate that classroom observers evaluated an item positively; crosses mean they evaluated it negatively. Thus, for example, no teacher could earn a positive evaluation of management combined with a negative evaluation of climate. For teachers of any skill, learning how to establish a safe learning climate precedes learning how to efficiently manage the classroom.

1.5 Theory behind the instruments: Proposed use

The proposed interpretation allows for both evaluative decisions (i.e., decisions regarding salary, tenure, or, in extreme cases, dismissal) and feedback. As outlined by Chapters 5 and 6, valid uses for performance evaluation require that the procedure meets specific conditions. If these conditions are not met, decisions about a teacher are unjustified. As outlined by Chapters 2, 3 and 4, the instruments can also be used to inform feedback. Item scores obtained with the instruments can be ordered cumulatively according to the predicted six stages. This is considered a unique aspect of these two instruments. Items included in the instruments are specifically selected on the basis of whether they fit the intended use of teacher feedback.

To use the outcomes described herein to provide feedback, it is crucial to confirm for each particular teacher, whether her or his development of teaching skill can be approximately described by six consecutive, cumulative stages. Researchers in the field of teacher development share varying viewpoints on whether it is valid to claim that all teachers develop similarly (an excellent overview of most recent theories on teacher development is provided by: Louws, 2016). Chapter 6 in specific examines the validity of this claim and provides some tools and statistics that can be used to trace individual teachers or lessons that deviate from the predicted ordering. This chapter is a response to the concerns expressed by scholars in the field of teacher development which have argued in favor of less hierarchical and more flexible interpretations of teacher development, because they were unconvinced that all teachers develop similarly (Berliner, 2001; Day, Sammons, Stobart, Kingston & Gu 2007; Huberman, 1993). They proposed instead that teachers' development can better be described using several nonhierarchical phases. Teachers can be grouped according to these phases, but teachers grouped in the same phase may develop very differently thereafter. The studies in this dissertation do not deny that an interpretation in terms of phases might offer a more accurate interpretation of teacher development; however, the studies do argue that an interpretation in terms of phases has modest practical utility. When applying phases to describe teacher development, it becomes impossible to advise teachers about specific steps regarding what to learn or develop next. As Richardson and Placier (2001, p. 913) put it, "the use of a very flexible approach to stages or phases may have taken us so far from the original concept of a stage theory that the usefulness of the work must be rethought." The studies in this dissertation present evidence that an interpretation in terms of cumulative stages is supported by empirical evidence and using it as such can be justified.

1.6 Types of evidence used: An introduction to the Rasch model

Most evidence in this dissertation is rooted in a specific type of statistical model, the Rasch model. It is part of a wider family of models commonly referred to as item response theory (IRT) models. To shed light on the importance of the evidence, a brief introduction to these statistical models is necessary. (For a more detailed introduction, see Bond and Fox 2007.)

As a statistical theory of measurement, IRT articulates sets of assumptions that must be verified before item scores can be interpreted and used validly (Bond & Fox, 2007, Fox, 2010). The Rasch model can validate whether data are cumulatively ordered, as exemplified by Figure 1.1 (Bond & Fox, 2007; Rasch, 1960). Cumulative order implies that some teaching practices are performed more frequently than other teaching practices and that the performance of more frequently observed teaching practices is required for performance of less frequently observed practices. The conditional argument distinguishes the Rasch model from other statistical models, and it provides clear definitions of some important concepts underlying any empirical investigation of differences in teaching skill, in particular the concepts of complexity, better, and improvement. First, the studies included in this dissertation interpret less frequently observed teaching practices as more complex. The term "complexity" does not refer to the practices themselves when studied in isolation, any behavior appears rather simple. Rather, it refers to the cumulative principle that teachers' use of more complex teaching practices requires them to perform these practices in parallel with less complex ones. Performing more practices at the same time makes teaching more complex. Second, valid evaluation of differences in teaching skill requires a clear definition of what constitutes better skill in teaching. The Rasch model can provide a clear definition of "better," because if the model fits, it follows that teachers obtaining higher evaluation scores have performed the same teaching practices as peers who have obtained lower evaluation scores, plus some additional teaching practices (see Figure 1.2). This guarantees that teachers evaluated as more successful are not using completely other or different practices in comparison to their less successful peers. Better teachers succeed in implementing additional teaching practices. Third, "improvement" refers to a teacher successfully adding a teaching practice, which suggests that the cumulative order can be applied to structure teachers' development and learning. Chapters 2 and 3 explain this interpretation in more detail.

Note that the aforementioned advantages are specific to the Rasch model approach[1] and that these interpretations are only valid if the assumptions of the Rasch model hold. Therefore, fitting the data to the Rasch model's assumptions is of fundamental importance and one of the main topics of investigation in this dissertation.

1.7 Structure of the dissertation

The dissertation is structured in eight chapters. Broadly, Chapters 2–4 discuss the measurement properties of the instruments. Herein, it is evaluated whether the teaching

[1] An exception is the non-parametric Mokken model, which can also used to test for cumulative item order (for details see: Meijer, 1994; Meijer, Sijtsma, and Smid, 1990).

practices included in the instruments can be ordered cumulatively and whether the observed order aligns with other theory concerning the development of teaching. Chapter 5 details how various evaluation procedures, which schools may pursue, may result in more or less reliable feedback or evaluative decisions. Chapter 6 turns to individual differences in teacher development and examines how to identify and act on the specific instance when a teacher does not show the expected order in teaching development. Chapter 7 summarizes the main conclusions. The final Chapter eight discusses limitations, alternative interpretations, consequences for use of the instruments and some directions for further research. The studies included in this dissertation address the following research questions:

1. 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)**
2. Can student questionnaire ratings of effective teaching practices be ordered cumulatively? And; How may the development of such a scale contribute to the knowledge about teacher development? **(Chapter 3)**
3. To what extent do observers and students agree on the cumulative ordering in teaching practice complexity? **(Chapter 4)**
4. 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)**
5. How many observed lessons show substantial deviation from the cumulative ordering? And; Do deviating lessons cluster with some particular teachers? **(Chapter 6)**

Chapter 2 elaborates on the internal structure and validity of the ICALT observation instrument to provide secondary school teachers with feedback. The sample used for this study contains 878 lesson observations at 119 schools across the Netherlands, of which 46.6% were from the Dutch inspectorate and the other 53.4% by peer colleagues. **Chapter 3** discusses the internal structure and validity of the “My Teacher” questionnaire using a sample of 1,590 questionnaires related to 68 teachers with varying experience (0–43 years) from one Dutch school. **Chapter 4** elaborates on the comparability between the

ICALT observations and the “My Teacher” questionnaire. The chapter presents evidence that students and classroom observers assign items similar interpretation and whether and how the ICALT observation instrument and “My Teacher” questionnaire can be merged into a single instrument. The sample contains 269 classroom observations and 2,876 student questionnaires evaluating the same 141 teachers. **Chapter 5** accentuates the importance of reliability of teacher evaluations, and it examines how the reliability of classroom observation increases if the number of lessons visited increases and the number of observers increases. This chapter is based on a sample of 198 lesson observations of 69 teachers by 62 observers obtained at eight schools. **Chapter 6** studies individual differences in the development of effective teaching and possible consequences for evaluation. If a teacher develops differently than would be predicted by the model, the evaluation approach would provide them inaccurate directions for improvement. This manuscript is based on the same sample as studied in Chapter 5. Finally, **Chapter 7** contains the main conclusions, and **Chapter 8** discusses some methodological limitations, and provides some recommendations for evaluation practice.

The **Appendix** contains an additional article published in *Pedagogische Studiën* regarding the reliability of teacher-assigned grades. It was originally intended to be part of this dissertation and is appended, because it presents some important insights regarding this original intention: i.e. to explore whether teacher-assigned grades can be used to assess teaching skill. If we could adequately diagnose teachers in need of assistance on the basis of teacher-assigned grades, schools could more efficiently target student questionnaire and classroom observation methods. However, the results showed that teacher-assigned grades are too unreliable and cannot be used for this purpose. Therefore, this idea was abandoned, and the data were used to explore whether report card grades were sufficiently reliable to make decisions about students, which is still a relevant issue but not part of the general research focus of this dissertation.

Chapter 2

Teacher evaluation on the basis of Classroom Observation

Abstract

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

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

2.1 Introduction

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

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

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

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

2.2 Background

2.2.1 Teacher development

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

2.2.2 Fuller's theory of teacher concerns

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

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

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

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

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

2.2.3 Teacher effectiveness literature

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

2.2.4 Integration of teacher development and teacher effectiveness literature

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

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

2.3 Method

2.3.1 Sample

The sample consisted of 958 teachers whose lessons were observed by trained observers in 119 schools located across the Netherlands. The observations were performed by either peers (53%) or inspectors from the Dutch inspectorate (47%). Teacher experience ranged from student teachers with 0 years of experience to those who had been teaching for 41 years. Of these teachers, 51% were men, and 25.7% held a master's degree. The sample included all education types, from preparatory secondary vocational education to university preparatory education, and students from all grades. The classroom subjects in which the observations took place were Dutch, history, math, biology, geography, English (as a foreign language), social science, science and physics, economics, French, German, philosophy, arts, drawing and construction, Spanish, Latin, music, and informatics. All observations took place between spring 2010 and summer 2011.

2.3.2 Instrument

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

2.3.3 Data selection procedures and missing data

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

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

2.3.4 Model

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

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

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

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

2.3.5 Data analysis and software

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

2.4 Results

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

2.4.1 Development sample

Parallel ICC. The Rasch model and two-parameter IRT model are nested models that differ only in that the latter allows for random item characteristic curves (ICC). Rizopoulos (2006) suggested comparing the fit of both models using the $\Delta\chi^2$ test. If the $\Delta\chi^2$ test is insignificant, the ICC's are approximately parallel. The results indicate that the Rasch model has slightly worse fit than the two-parameter IRT model ($\Delta\chi^2 = 45.14$, $df = 31$, $p < .05$). Closer inspection of the random slope parameters the (a-parameters) reveals that item slopes varied from 1.30 (SE = .23) < a < 2.46 (SE = .42). These slopes do not statistically deviate (±1.96*SE) from the average slope ($M(a)$ = 1.75). The only exception is item 22, "explains the lesson objectives at the start of the lesson," for which the ICC slope (a) = 1.05, SE = .18. When deleting item 22, the Rasch model and two-parameter IRT model have identical fit ($\Delta\chi^2$ = 36.54, *df* = 30, *p* = .19). Therefore, all items other than item 22 exhibited approximately parallel ICC.

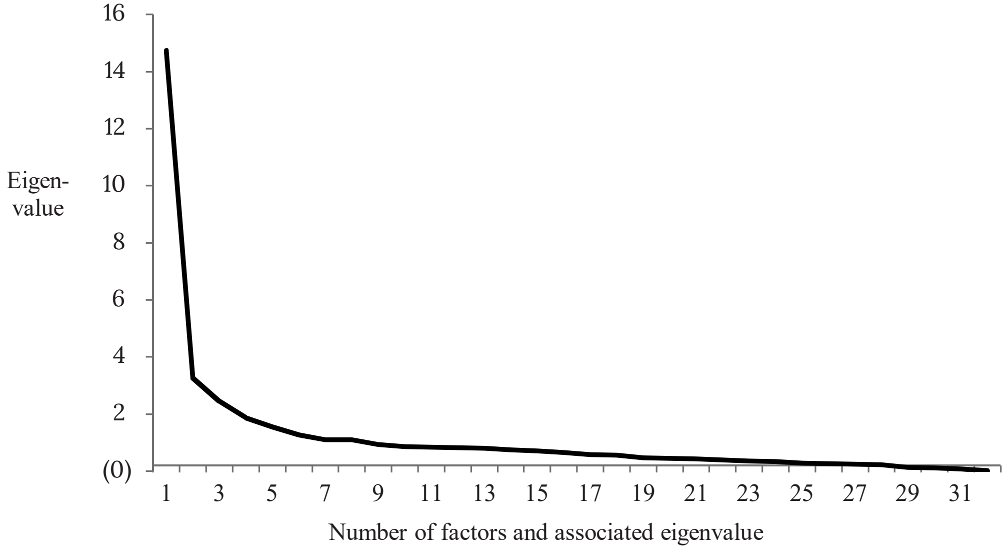
One-dimensionality. The assumption of one-dimensionality is difficult to (dis)confirm. Despite the fact that many tests have been proposed to evaluate one-dimensionality (e.g., Haberman, 2008; Stout, 1990; Timmerman, Lorenzo-Seva, & Ceulemans, in press), there is not much consensus about any statistical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

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

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

Figure 2.1

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



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

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

Ponocny's T_1 and T_{1m}. Rasch (1960) originally proposed but never completed a nonparametric test to assess model fit. Ponocny (2001) based his family of *T*-statistics on Rasch's original intentions, using the raw sum scores for the items and persons. From the observed sum scores, this test generates alternative matrices with identical sum scores, then tests whether these alternative matrices all fit the Rasch model. The test of local independence uses the number of extreme scoring patterns: {11} or {00} and tests whether the number of extreme scoring patterns on two items *j* and *k* is higher (in case of T_1) or lower (in case of T_{1m}) than the number of extreme scoring patterns expected by the Rasch model (see Koller & Hatzinger, 2013; Ponocny, 2001). With an instrument of 32 items, the tests evaluate violations of local independence for 496 item pairs. With so many tests, some violations may occur simply due to chance (e.g., Koller and Hatzinger, 2013; Ponocny, 2014, personal communication). Koller and Hatzinger (2013) therefore propose correcting for chance inflation by dividing alpha by the number of item pairs tested. With this correction, the alpha level becomes (.05/496) = .0001.

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

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

Next, the T_1 tests reveal any positive increasing residual correlations. Positive residual correlations indicate that thee teaching practices described by two items share an additional characteristic, other than teaching skill, due to which item scores are more similar than can be explained by teaching skill alone. This test diagnosed six item pairs (we list their domains in parentheses): item 3, "supports student self-confidence" (safe learning climate) with item 17, "boosts the self-confidence of weak students" (student activation); item 21, "provides interactive instruction" (student activation) with item 31, "encourages students to think critically" (teaching learning strategies); item 22, "explains the lesson objectives at the start of the lesson" (student activation) with item 23, "checks whether the lesson objectives have been achieved" (differentiation); item 24, "offers weak students additional learning and instruction time" (differentiation) with item 25, "adapts processing of subject matter to student differences" (differentiation); item 27, "teaches students how to simplify complex problems" (teaching learning strategies) with item 32, "asks students to reflect on approach strategies" (teaching learning strategies); and finally, item 28, "encourages the use of checking activities" (teaching learning strategies) with item 29, "teaches students to check solutions" (teaching learning strategies).

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

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

2.4.2 Summary of main findings for the development sample

In the development sample, only item 22, "explains the lesson objectives at the start of the lesson," exhibited misfit with the cumulative stagewise pattern. It has a deviating ICC and a low factor loading, and one test (Ponocny's T_{1m} and T_1) confirmed that it violates local independence. We note further that item 22 previously has been shown to violate model assumptions (see for example: Van de Grift, Helms-Lorenz, & Maulana, 2014).

2.4.3 Cross-validation

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

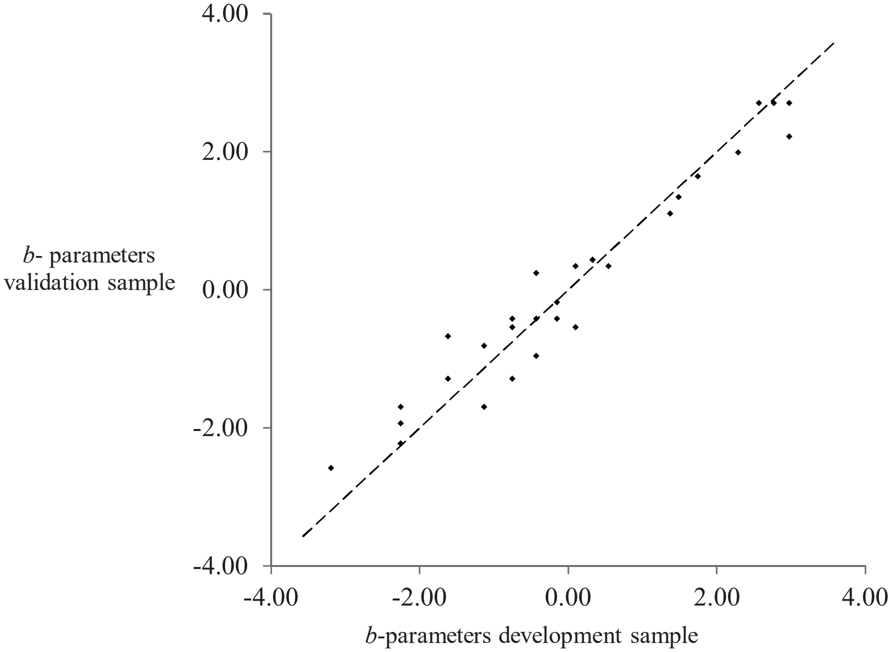
We reassessed the assumption of one-dimensionality using a CFA. The model fit again is mixed with RMSEA below the .05 threshold, but CFI and TLI above the threshold (χ^2 (465, n = 439) = 906.69, p = .00; CFI = .94, TLI = .94; RMSEA = .047 [90% CI = .042 - .051]). The scree plot again showed one dominant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

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

This cross-validation accordingly confirmed that all items except item 22 fit the invariant cumulative and one-dimensional ordering. We recommend that the item should be discarded from the instrument when a Rasch analysis is applied. Tests for local independence consistently diagnosed some underlying patterns that might help clarify what creates multidimensionality in the current measures of teaching skill. In particular, the negative residual correlations between effective teaching practices in the domain of efficient classroom management and those in the domain of differentiation request further exploration.

Figure 2.2

The goodness-of-fit (GoF) plot.



Notes. The 31 dots represent the 31 items. The x-axis gives the item complexity (*b*-) parameters for the development sample. The y-axis gives the item complexity parameters for the validation sample. The dashed line represents complete invariance, and deviations from the dashed line indicate deviations from sample invariance.

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

2.4.4 Comparison of cumulative ordering with Fuller's theory

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

Table 2.1

Final cumulative ordering in effective teaching practices

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

--- continues next page ---

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

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

The results in Table 2.1 also show that the most complex practices in less complex domains surpass the least complex practices of more complex domains. This pattern suggests that teachers do not develop all the skills in one domain first, before proceeding to the next domain. Rather, the transition from one stage to the next is gradual. Some domains, such as efficient classroom management and quality of instruction, appear almost equally complex, which suggests that they might develop simultaneously.

2.5 Conclusion

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

2.5.1 Limitations

The sample included 958 teachers working in 119 schools. Technically, the data should be considered nested, with teachers nested in schools. While the parameters are estimated using multilevel Rasch model techniques, the assumptions tests have not been corrected for the multilevel structure. We checked whether the item parameters estimated by the assumption tests differed from the parameters estimated in the multilevel Rasch model, to verify validity of the assumption tests. No large deviations were found between them. However, the standard errors of item parameters were larger in the multilevel Rasch model, compared to the standard errors estimated by the assumption tests. This implies that by not correcting for the nested data structure in our assumption tests we plausibly are overly strict and wrongly removed items that actually did fit. We note that assumption tests to evaluate fit of Rasch models in nested datasets are still at a developmental phase (e.g., de Boeck et al., 2011; Fox, 2010) and have not yet been incorporated in standard software packages. We consider our approach as the best option currently available.

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

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

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

Chapter 3

Teacher evaluation on the basis of student ratings

Abstract

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

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

3.1 Introduction

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

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

3.2 Theoretical background

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

3.2.1 Theories of teacher development

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

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

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

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

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

3.2.2 Teacher effectiveness literature and development in teaching skill

Several reviews and meta-analyses address the relation between teaching practices and student achievement (Hattie, 2009; Kyriakides, 2013; Marzano, 2003), and though they use different labels, they show consistently that similar categories of teaching practices enhance student achievement. We consider six broad domains of teaching practices that can be observed within classrooms: creating a safe learning climate, efficient classroom management, quality of instruction, student activation, teaching learning strategies, and differentiation (the questionnaire items are included in Appendix C [English translation] and E [Dutch version]). Van de Grift (2014) provides an extensive literature review to account for the six domains. In addition, Maulana, Helms-Lorenz, & Van de Grift (2015) describe connections between these six domains and the classroom assessment scoring system (CLASS) and the framework for teaching (FFT) observation protocol, both of which are currently employed in the Measures of Effective Teaching (MET) project. They conclude that the six domains coincide with all the clusters of the FFT and CLASS.

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

Table 3.1.

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

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

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

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

3.2.3 Evaluation method: student ratings

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

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

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

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

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

3.3 Method

3.3.1 Sample

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

3.3.2 Measurement instrument

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

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

3.3.3 Design and missing values

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

3.3.4 Cross-validation procedure

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

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

In total, 1,590 students are included in the development and validation samples. The other 672 students were omitted. Subsamples did not differ in student total test scores (F(2, 2214.58) = .27, p = .79) or in student age (F(1, 2193.89) = .20, p = .66), though they did differ slightly on student gender ($\chi^2(2, n = 2{,}226) = 6.90$, $p = .03$). The omitted sample had 57.8% girls, while the two randomly selected samples; 51.3% and 52.4%.

3.3.5 Model specification

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

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

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

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

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

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

3.3.6 Data analysis

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

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

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

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

3.3.7 Software

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

3.4 Results

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

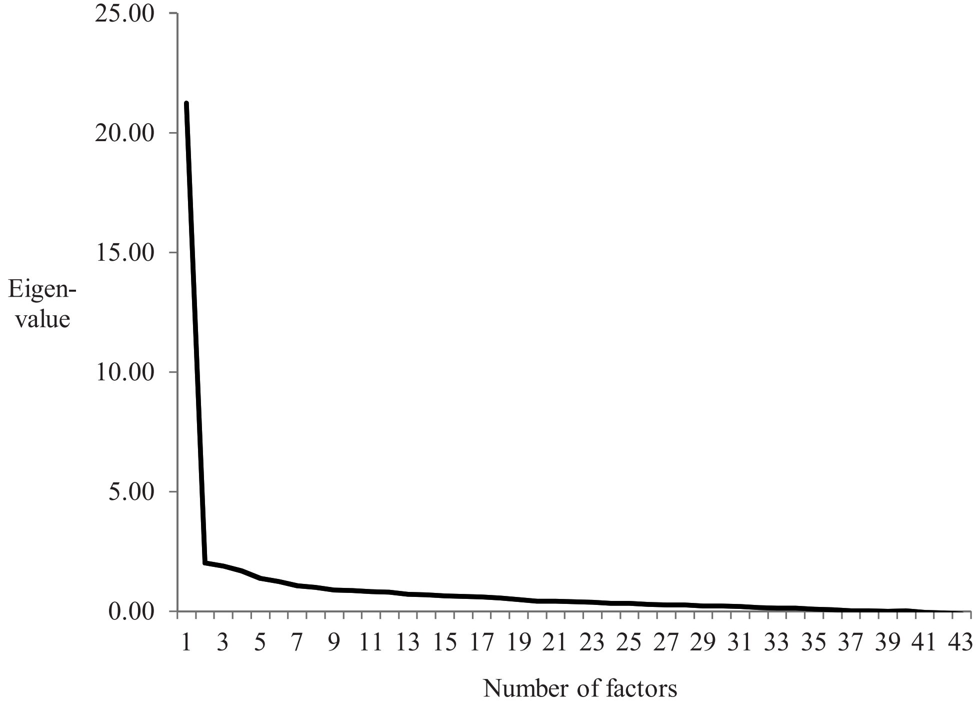
3.4.1 Development sample

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

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

Figure 3.1

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



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

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

Chance should have an important position in evaluating the *T*-statistics results (I. Ponocny, personal correspondence, September 30, 2014). The *T*-statistics pair every item with 42 other items. Therefore, a criterion of two violations per item would reflect an alpha criterion of .05. However, their considerable power together with the slight overlap in item content (both within and between domains) and students’ differential grammar ability, makes that some additional violations are almost inescapable and may be tolerated. On the basis of these considerations we decided to set a more lenient criterion of 5 violations.

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

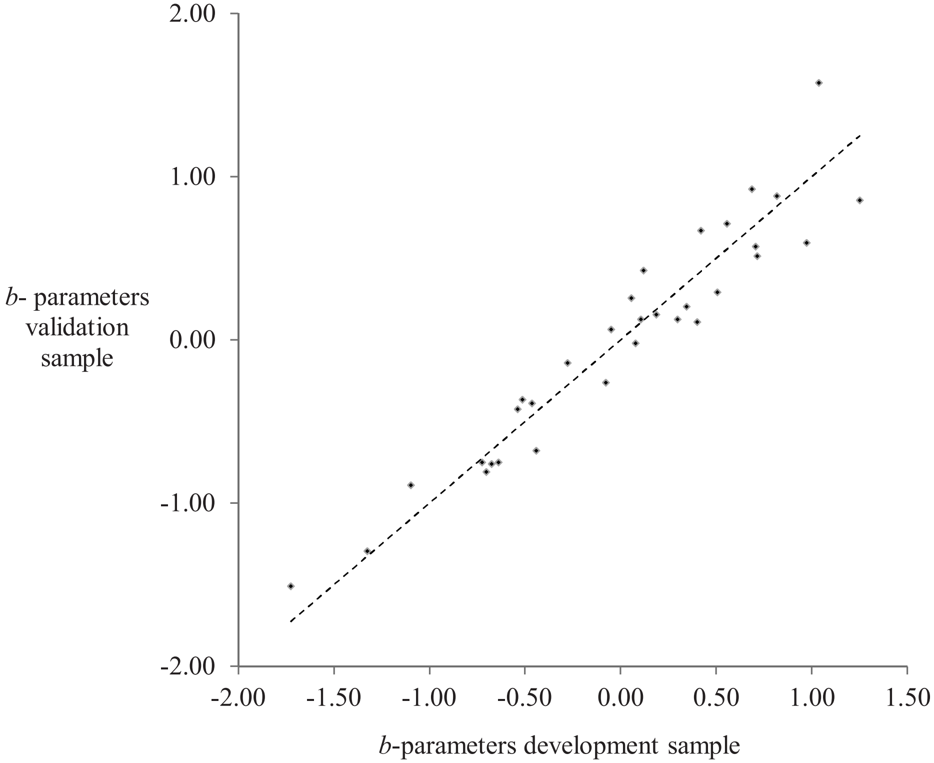
3.4.2 Validation sample

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

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

Figure 3.2

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



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

3.4.3 Final questionnaire

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

Table 3.2

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

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

--- continues next page ---

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

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

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

Broadly, the cumulative ordering in Table 3.2 aligns with descriptions of teacher development: It starts with teaching practices that establish a safe learning climate and quality of instruction and ends with teaching practices associated with differentiation and teaching learning strategies. This result confirms our predicted cumulative ordering in complexity. Furthermore, the ordering in Table 3.2 shows considerable within-domain variation, for efficient classroom management in particular. This result suggests that the least complex skills of (more complex) domains may precede the development of the most complex skills of other (less complex) domains. This finding fits with discussions about the limitations of perceiving teacher development in rigid stages, which have continually suggested that descriptions (and measurement) of development in invariant stages is inappropriate and that more flexibility is desirable. By establishing the cumulative ordering at the level of teaching practices, the instrument avoids the requirement of a complete invariant hierarchical ordering in domains. We also note that the ordering includes teaching practices from all six previously identified domains of effective teaching. Our strict procedures for selecting teaching practices thus did not exclude any domain from the instrument; omitting 27 items describing various teaching practices seemingly did not produce any unacceptable loss of information. In support of this assertion, we computed the correlations of the evaluation scores for teaching skill measured with the original 59 items versus those measured by the 32 selected items. A high correlation would suggest that excluding the 27 items had a minor impact on final evaluations of teaching skill. Indeed, we find that the Pearson product moment correlation between teacher skill scores obtained from the 59- versus 26-item instrument was $r = .99$, with $n = 84$ and $p < .00$.

3.4.4 Measurement reliability

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

Table 3.3

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

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

--- continues next page ---

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

3.5 Conclusion

Our results confirm the main premise: Effective teaching practices can be ordered cumulatively, from basic to more complex. Broadly, the cumulative ordering observed is in accordance with Fuller's (1969) theory on teacher development, which states that teachers are first concerned with the self, then with the task, and finally with their impact on student learning. Furthermore, the cumulative ordering mirrors the ordering found on the basis of classroom observations (e.g., van de Grift, et al., 2014; van der Lans, et al., 2017). Thereby, the validation of a cumulative ordering also provides some initial insights in the development of effective teaching practice. These findings represent an important step toward instruments that can provide truly formative feedback. In the future, the instrument developed here could provide an alternative to those in use currently, which can score teachers' current skill but lack the underlying, empirically validated, cumulative ordering required to present objective advice about the next steps to improve.

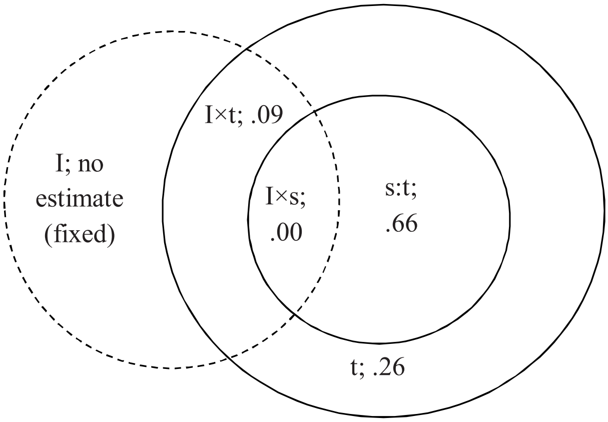
3.5.2 Limitations

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

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

Figure 3.3

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



Chapter 4

Combining student ratings and classroom observations in teacher evaluation

Abstract

Using item response theory (IRT) this study explores whether items of a student questionnaire and a classroom observation instrument can be viewed as measuring one latent construct, namely teaching skill. The data comprises 269 lessons of 141 teachers which were evaluated using the international comparative analysis of learning and teaching (ICALT) observation instrument and the “My Teacher” student questionnaire. Rasch model analysis confirms that items from both instruments fit a one-dimensional cumulative ordering. Also, students and observers are found to interpret items measuring similar teaching practices equally. However, students and observers can still disagree on which teaching practices any particular teacher uses. After removal of biased items, the correlation between student ratings and classroom observations remains moderate (r = .34).

Based on the manuscript: Van der Lans, R. M., Van de Grift, W., J., C., M., & Van Veen, K. (2016). *Synthesizing teacher evaluation methods: The case of student ratings and classroom observations of teaching*. Manuscript submitted for publication.

4.1 Introduction

This study is motivated by a practical problem: specifically, student ratings and classroom observations provide different evaluations to teachers (e.g., Feldman, 1989; Maulana & Helms-Lorenz, 2016), yet both evaluation methods are considered valid and reliable (e.g., Benton & Cashin, 2012; Kane et al., 2012). This situation considerably complicates the evaluation practice: Although research communicates that schools, districts, and states can choose between these evaluation methods to provide teachers with reliable and valid evaluations, the same research indicate that teachers receiving poor evaluations from one method might have received good evaluations from the other method.

This contradictory situation is partially due to a research tradition in which investigators study the psychometric properties of instruments in isolation of other instruments designed to measure the same latent construct. As a result, these instruments may show high reliability individually, though, when compared with other instruments, show considerable diversity. To reduce this diversity, some researchers advise averaging multiple evaluation methods into a single composite scores (e.g., Mihaly, McCaffrey, Staiger & Lockwood, 2013; Peterson, 2000; NCTQ, 2013). The logic behind this advice is that if both methods measure the same construct, then the average composite score should be considered more reliable and less susceptible to specific biases associated with each particular method. However, because it is unclear whether student questionnaire and classroom observation methods measure the same latent construct, it is unclear whether this composite actually represents one latent variable or constitutes a mix of two distinct constructs.

Researchers put forth various theoretical explanations for the different evaluations between students and classroom observers. Kunter and Baumert (2006) and Maulana and Helms-Lorenz (2016) propose that different evaluation methods will show overlap but that each method also measures specific and unique elements of the construct: teaching skill. These authors maintain that any attempt to synthesize evaluation methods will remove the valuable information that makes each method unique and thus are not worth pursuing. Others question students' expertise to evaluate teaching (e.g., Peterson, 2000), especially younger students and students in lower educational tracks, who may lack the reading and comprehension skills necessary to provide valid item responses. In this line of thinking, classroom observers are trained experts and thus provide more valid indications about teaching than students, who are untrained novices. Further complicating these discussions, virtually no information about the differences in interpretation between classroom observers and students is available at the item level (see also Maulana & Helms-Lorenz, 2016). Therefore, currently we can only speculate about whether the varying evaluation outcomes reflect differences in interpretation at the item level.

In addition, we note the issue of each method's cost-efficiency. Student questionnaires are more cost-efficient than classroom observation, which makes them an attractive option to replace classroom observation (e.g., Van der Lans, Van de Grift, & Van Veen, 2015; Keuning, Van Geel, Visscher, & Fox, 2016). However, when different methods provide different evaluations to teachers for reasons not completely understood, such decisions may have unintended consequences.

In this study, we use item response theory (IRT) to explore similarities in item responses between students and classroom observers responding to items that measure the same latent variable: teaching skill. Our aim is to explore the factors that underlie the disagreement between students and classroom observers and verify whether they can be attributed to differences in the interpretation of the items and, thus, the construct.

4.2 Background

4.2.1 Context and purpose of the instruments

Current educational policies put increasing emphasis on teacher evaluation and assessment (e.g., Looney, 2011; Mourshed, Chijioke, & Barber, 2010). This study took place within the Netherlands, in which also a growing need has arisen for schools to evaluate and reward effective teaching and to give teachers specific advice on how to improve their teaching (Organisation for Economic Co-operation and Development [OECD], 2016).

The two instruments investigated in this study were developed in this context. The observation method is the international comparative analysis of learning and teaching (ICALT), and the student rating instrument is the "My Teacher" questionnaire. The two instruments have been in use for several years and in various projects, including a national teacher induction project, a regional project directed at improving teacher evaluation at low-performing schools (Van de Grift, 2014), and multiple global projects in which the instrument properties are compared across various countries. Using data from these projects, the reliability and validity of evaluative decisions and feedback based on the two instruments has been investigated thoroughly (e.g., Van de Grift, Helms-Lorenz, & Maulana, 2014, Van der Lans, Van de Grift & Van Veen, 2016, Van de Grift, 2014).

Specific attention has been given to the provision of feedback. To improve teacher feedback, these works connect teaching effectiveness literature with theory on teacher development to arrive at a conceptual understanding how teachers develop skill in and may learn effective teaching practices. This understanding is fundamental to our current aims and is briefly elaborated upon.

4.2.2 The conceptual understanding of how the instruments evaluate teaching skill

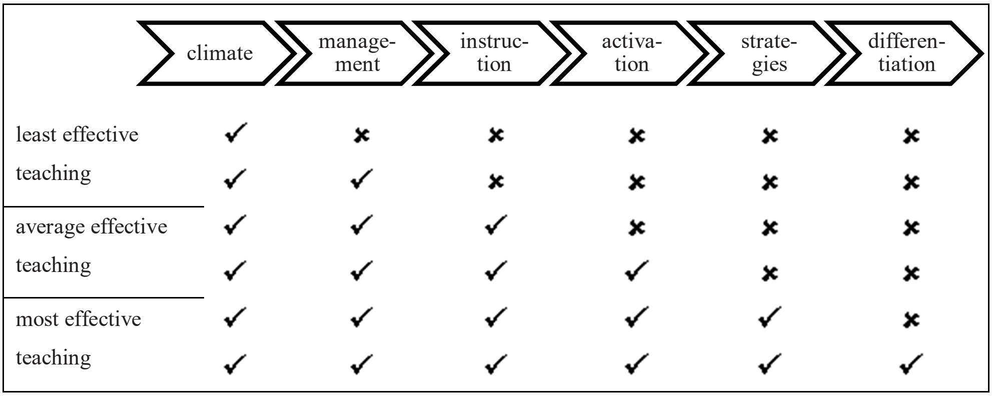
An important assumption behind these evaluation policies is that when provided with a reliable identification of a teacher's skill in teaching, evaluators can provide valid and specific feedback regarding what to improve. In this context, if teachers receive no clues about what to improve, their performance evaluation provides little perspective. As Firestone (2014) argues, this situation could even have detrimental effects on education in that it may demotivate teachers. Therefore, it is important that evaluation instruments clarify how they define and conceptualize teaching skill and how they relate current teaching skill to specific advice on how to improve current teaching.

We conceptualize teaching skill as effectiveness. Effective teachers have a large repertoire of teaching methods, behaviors, and strategies (hereinafter, simply "teaching practices") that positively relate to student achievement, such as the practices mentioned in the work of Muijs et al. (2014), Marzano (2003), and Strong (2011). Because the instruments are grounded in literature on teacher effectiveness, the items in the instruments show considerable overlap with items mentioned in other classroom observation and student rating instruments, including the Tripod (e.g., Bill & Melinda Gates Foundation, 2012), classroom assessment scoring system (CLASS) (e.g., Pianta & Hamre, 2009), and the framework for teaching (FFT) (e.g., Danielson, 2013).

To conceptualize improvement, we rely on stages, or cumulative development. In this conceptualization, teachers improve their teaching skills when they successfully add a practice to their repertoire. Moreover, if improvement in teaching skill is conceptualized as cumulative, it logically follows that less complex practices must precede more complex ones and that success in adding a practice will depend on whether its complexity is on par with the teacher's specific skill. In line with Fuller's (1969) three-stage theory of teacher development, we hypothesize a hierarchical and cumulative development in effective teaching practices (for details, see Van der Lans, Van de Grift, & Van Veen [2015, 2016], who discuss how the proposed cumulative ordering fits other work on teacher development, including Berliner [2001] and Huberman [1993]). Teaching practices involving classroom climate and respectful relationships are the least complex to develop, and competence in them is a prerequisite condition to competence in moderately complex practices such as classroom management and basic instruction. Competence in moderately complex practices, in turn, is a prerequisite to more advanced, complex teaching practices, including interactive instruction, teaching learning strategies, and differentiation. Previous work (Van der Lans, Van de Grift, & Van Veen, 2015, 2016) further refines these three stages into six cumulatively ordered domains (see Figure 4.1) (For a detailed description of the domains, see Van de Grift 2014.)

Figure 4.1

Hypothesized cumulative ordering in domains of teaching practices. Check marks reflect positive observations, and crosses signify negative observations.



	climate	management	instruction	activation	strategies	differentiation
least effective teaching	✓	✗	✗	✗	✗	✗
	✓	✓	✗	✗	✗	✗
average effective teaching	✓	✓	✓	✗	✗	✗
	✓	✓	✓	✓	✗	✗
most effective teaching	✓	✓	✓	✓	✓	✗
	✓	✓	✓	✓	✓	✓

Previous results about the cumulative ordering show considerable consistency across these works, but results regarding the ordering of the two most complex domains (i.e., learning strategies and instruction differentiation) is mixed. Specifically, research using the classroom observation method in primary education shows learning strategies to be the most complex (e.g., Van de Grift, Van der Wal & Torenbeek, 2011), a finding corroborated by research using the student questionnaire “My Teacher” in secondary education (e.g., Van der Lans, Van de Grift, & Van Veen, 2015). However, research using the classroom observation method in secondary education shows differentiation to be the most complex (e.g., Van de Grift, Helms-Lorenz, & Maulana, 2014).

4.2.3. In search of complementary evaluations

Cumulative ordering proves valuable feedback to teachers, such that they can understand what they have achieved already, what they should learn (which we refer to as the teacher's "zone of proximal development"), and what is yet too complicated to implement. However, on occasion evaluators find that students give other indications about the teachers' current teaching skill—and thus teachers' zone of proximal development—than observers. The disagreement between classroom observers and students is confirmed by other empirical works. Howard, Conway, and Maxwell (1985) report a correlation of r = .24 between classroom observers' scores for a one-time lesson visit and student ratings. Similarly, De Jong and Westerhof (2001) report an average correlation of r = .12. In recent work using the same classroom observation instrument and student questionnaire used in this study, Maulana and Helms-Lorenz (2016) report that the correlation between scores from classroom observers' one-time lesson visits and student ratings is r = .26. Again, this correlation suggests little overlap, leading the authors to question whether students and observers measure the same construct.

4.3.4 This study

This study applies IRT to investigate in more detail what underlies the disagreement between students' and classroom observers' ratings. Given the low correlations found in previous studies, it seems unrealistic to expect agreement between students and classroom observers on the competency of any particular teacher. However, this disagreement does not necessarily imply that they disagree on their interpretation of the items. Students and observers may order items describing equal teaching practices similarly and assign these similar complexity. In this particular case students and classroom observers agree on the complexity of teaching practices and the low correlation reflects disagreement about which practices the teacher uses. The latter might be explained by the different information that students and classroom observers can access. Students have experienced all the lessons with a teacher and their rating generalize across many lessons, whereas classroom observers have only a snapshot; if teaching skill varies from lesson to lesson, the observer may have an overly positive or negative snapshot and therefore position the teacher accordingly, decreasing the correlation with student data. Thus, the focal research question is as follows:
To what extent do observers and students agree on the cumulative ordering in teaching practice complexity?

4.3 Method

4.3.1 Data

Data is selected from three different research projects in the Netherlands. The first is an independent research project focused on the evaluation of in-service teachers working at 13 schools located across the Netherlands. The second is a research project, funded by the Dutch ministry of education, and is located in the Northern provinces in the Netherlands. It focuses on the implementation of teacher evaluation in 11 weak performing schools as judged by the Dutch inspectorate of education. The third project is a ministry financed project focused on evaluation and improvement of beginning teachers (≤ 3 years of experience).

The sample comprises 269 classroom observations of 141teachers having varying experience (0-40 years). The 141 teachers are evaluated by 1,237 students of which 46.3% is male and student age ranged between 11 and 18 years ($Mdn_{(\text{age})}$ = 14 years). All types of education are included including preparatory secondary vocational education, preparatory higher vocational education, and university preparatory education. Class size varied from 5 students (in vocational education) to 30 students (class-size mode is 24 students). The same 141 teachers are also evaluated by 93 observers with varying range in teaching experience (0-40 years). All observers are trained. Inter-rater agreement varied between schools and research projects but all above 70%.

4.3.2 Instruments

“My Teacher” student questionnaire. The “My Teacher” questionnaire has calibrated and validated in two previous works (Maulana et al., 2015; Van der Lans et al., 2015). The study by Maulana had a particular focus on beginning teachers in secondary education. The study by Van der Lans had a particular focus on in-service teachers. The complete questionnaire counts 40 items, some selected by Maulana et al. others by Van der Lans et al. Because the current sample includes in-service teachers having various years of experience, the subsample of 28 items previously identified by Van der Lans, et al. (2015) is used for this study. Each item reflected a statement related to the teacher’s teaching practices. Example items are: “my teacher applies clear rules”, “my teacher stimulates my thinking”, and “my teacher ensures that I use my time effectively.” Items can be grouped in six domains: safe learning climate (SLC), efficient classroom management (ECM), quality of instruction (QOI), activating teaching methods (ATM), teaching learning strategies

(TLS), and differentiation in instruction (DII). An extensive review accounting for these six domains is presented by Van de Grift (2014). Students rate items on a dichotomous scale coded “0” = rarely and “1” = often.

The international comparative analysis of learning and teaching (ICALT) observation instrument. A subsample of 31 items of the ICALT observation instrument is used (complete instrument counts 32 items). The selection is again based on previous work which indicates that these 31 items fit the cumulative ordering (Van der Lans, Van de Grift, & Van Veen, 2016). An example item is “uses teaching methods that activate students.” Items can be grouped in the same six domains SLC, ECM, QOI, ATM, TLS, and DII. Observers score the items using 4-point scale: 1 = not performed, 2 = insufficiently performed, 3 = sufficiently performed and 4 = well performed. To make comparison possible the, original coding 1 and 2 are recoded 0 and, the original coding 3 and 4 are recoded 1.

4.3.3 Data preparation, cross validation and missing values

In the complete dataset, each classroom observation of teacher “A” is accompanied by an entire class of student ratings. This provides a complex dataset, which is modeled using multilevel Rasch model approach (e.g., de Boeck et al., 2011; Doran, Bates, Blies & Dowling, 2007). However, fit tests for the multilevel Rasch model currently are not implemented in standard statistical software and, therefore, we need to rely on the regular Rasch model to assess model fit. To assess model fit, we randomly selected one student out of each class and connected them with their corresponding classroom observation. This resulted in a dataset comprising of 59 items, i.e. the sum of the number of items comprising the classroom observation instrument (=31) plus the number of items comprising the student questionnaire (=28).

A cross-validation procedure is used to check whether the results are not based on an accidental random selection of students. We randomly selected a second group of students to construct a second sample. This validation sample contained the same classroom observations but in combination with a different set of students.

Development sample. Some classroom observations have missing values on more than one-third of the 31 item responses (n = 10). These are discarded from the analysis. Furthermore, three classroom observations count less than two valid item responses within one of the six domains and are also discarded (n = 3). Using these same criteria all selected student questionnaires were eligible. After removal of these 13 cases, the sample counted 256 classroom observations corresponding to 256 student ratings. These 256 cases counted 120 missing values, which is 0.8% of all 15,104 item responses.

Validation sample. The same classroom observations are included in the validation sample. Again the 10 observations counting more than one third missing item responses and the three observations counting less than two valid item responses within a domain are discarded. Again all student questionnaires were eligible. The validation sample counted 256 classroom observations connected with 256 student ratings. These 256 cases counted 131 missing values, which is .9% of the 15,104 item responses.

4.3.4 Design and analysis plan

The analysis first concentrates on the ordering of items. Previous research shows that the items of the "My Teacher" questionnaire and the ICALT observation each fit Rasch model assumptions. In this study, it is examined whether they together also fit the Rasch model assumptions. If the Rasch model fits, this would provide evidence that disagreement between observers and students does not reflect that they measure different constructs.

4.3.5 Models and software

As a first step, we analyzed whether the items of different instruments together fit the Rasch model assumptions. The Rasch model specifies three model assumptions. The three assumptions are: local independence, one-dimensionality, and parallel item characteristic curves. Local independence implies that item residuals are uncorrelated. In this study, local independence is assessed using Ponocny's (2001) T_1 and T_{1m} statistics. These are included in the R package eRm (Mair & Hatzinger, 2007). The study examines one-dimensionality as the consistency of an item's complexity (or b-parameters) across random subgroups (Andersen, 1973). To evaluate this consistency, the original sample will be randomly split ten times in two equal halves. Using the Andersen (1973) log-likelihood ratio (LR)-test, we evaluated whether for each random split item complexity in both subgroups remains similar. Finally, parallel item characteristic curves (ICC) evaluate whether item complexity remains similar between varying levels of teaching skill. The Andersen (1973) Log-likelihood ratio test is used again, but now the sample is split using the median teacher evaluation total score.

In the subsequent step, the items that fit the cumulative one-dimensional ordering are selected. The complete sample is used including all student questionnaires. Using the R package lme4 (Bates et al., 2014), a multilevel Rasch model is specified (e.g., de Boeck et al., 2011) to estimate the complexity of the item parameters of the student questionnaires and the classroom observation instrument. We specified a random effect for observer—which could either be a student nested in a class or an observer—, a random effect for teacher and a random effect for item. The R package arm (Gelman et al., 2015) is used to generate the item standard errors.

4.4 Results

Results are reported in three subsequent steps. First Rasch model fit with the three assumptions is evaluated in the development sample. In this step we report about items that misfit the model and which need to be discarded. Then, model fit to the remaining item set is reevaluated in the validation sample. No further item selection is attempted and the results focus is on whether the selected items again fit the Rasch model assumptions. In the third step, the cumulative ordering is reported and discussion focuses on whether the students and classroom observers evaluate practices classified as belonging to the same domain as being approximately similar in complexity. Finally, we turn to the correlation between the student questionnaire and classroom observation method to verify whether resolving measurement bias increases the correlation between methods.

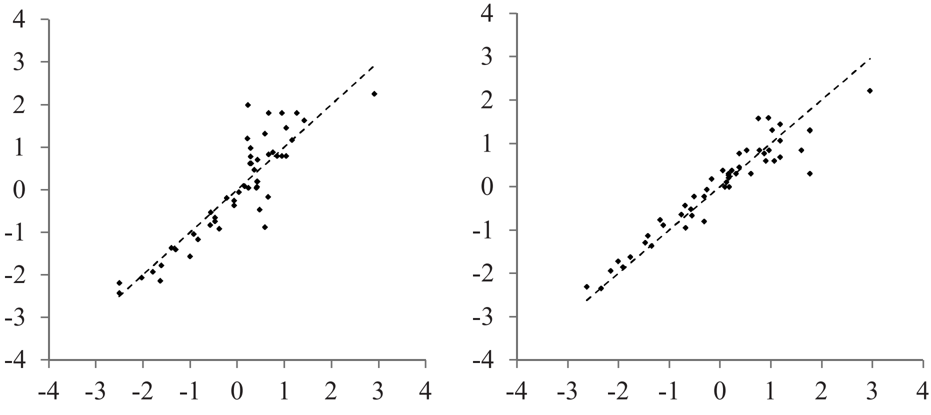
4.4.1 Evaluating model assumptions

Local independence. Ponocny's T_{1m} statistic diagnoses two "My Teacher" questionnaire items as showing more than one negative residual correlation, in specific: S27 "Teaches me to summarize," and S28 "Explains how I should study something." The negative residual correlations all involve pairings with ICALT items within the domains activating teaching methods (ATM) and teaching learning strategies (TLS). To improve model fit, these two items are discarded. Ponocny's T_1 statistic, identifies 27 positive residual correlations. Two broad patterns are evident. First, residual correlations all involve pairings of items from the same method: i.e., student-student or observer-observer. Secondly, the number of positive residual correlations is greater for the observation instrument and these mostly involve items within the domains: differentiation in instruction (DII) and TLS. After removing 7 items, the remaining 50 items show two decreasing residual correlations and <10 increasing residual correlations. The list is considered to be sufficiently locally independent. Removing these items did not seem to result in an unacceptable loss of information. Both instruments still cover all six domains.

One-dimensionality. Using a random number algorithm, the sample was ten times randomly split in two. Andersen (1973) Log-likelihood Ratio-tests shows significant deviations. Test values range from $\chi^2(df = 49) = 38.75$, $p = .85$ to $\chi^2(df = 49) = 55.72$, $p = .24$. This suggests that the items have approximately similar cumulative ordering for any random selection of teachers. Using a Goodness-of-Fit (GoF) plot, Figure 4.2 graphically portrays the consistency in item ordering. In a GoF-plot the item ordering of one subsample is plotted against the ordering in the other subsample. The solid line presents the item b-parameters in the first subsample, and the dots represent the b-parameters in the other subsample. The distance of each dot to the solid line indicates the difference in item b-parameters between the two subsamples.

Figuur 4.2

The Goodness of Fit (GoF) plot fo the least (left) and best (right) fitting subgroups.



Parallel ICC. To test the assumption of parallel ICC, the Andersen Log-likelihood Ratio-test (LR-test) (1973) is used. Here, the LR-test examines whether item complexity is approximately similar for teachers evaluated as having above average teaching skill and teachers evaluated as having below average teaching skill. The test included 50 items. Test results suggest that items approximately have parallel ICC ($\chi^2(df = 49) = 66.26$, $p = .051$).

4.4.2 Validation of Rasch model assumptions

The findings in the development sample are reassessed in the validation sample. Ponocny's T_{1m} diagnosed five item pairs violating local independence due to negative residual correlations. Two items; O32 "asks students to reflect on approach strategies" and O17 "boosts the self-confidence of weak students" counted more than one violation. These two items also had been diagnosed in the development sample but these were then considered acceptable. Based on this additional information, we decided to remove these two items and continue with the remaining 48 items. The 48 items counted one negative residual correlation. The T_1 statistic diagnosed 10 item pairs violating local independence due to positive residual correlations.

One-dimensionality – in terms of consistency in item ordering – is not violated. The Andersen LR-test values range from $\chi^2(df = 47) = 29.81$, p = .98 to $\chi^2(df = 47) = 63.36$, p = .06. Also, the Andersen LR-test showed no violations of parallel ICC assumption ($\chi^2(df = 45$[2]$) = 47.29$, $p = .38$. In sum, beside these few violations of local independence this set of items is found to measure an identical construct.

4.4.3 The progressive development in teaching skill

The Table 4.1 shows the established cumulative item ordering. The domains are abbreviated: safe learning climate (SLC), efficient classroom management (ECM), quality of instruction (QOF), activating teaching methods (ATM), teaching learning strategies (TLS), and differentiation in instruction (DII). The Table shows how the predicted ordering evolves in comparable pace among the two instruments and that items in both evaluation methods cover all six domains. This further adds to the validity of the "My Teacher" questionnaire and ICALT observation.

The comparability between classroom observation items and student questionnaire items is sometimes striking. For example, the classroom observer item O4 "ensures mutual respect" ($b = -.77$) and the student questionnaire item S8 "ensures that I treat others with respect" ($b = -.76$) receive almost identical item parameters, suggesting that observers and students agree about the complexity of this aspect of teaching. Also, the list provides information about how students interpret items. For example, the item S40 "helps me if I do not understand" is assigned similar complexity as the item O3 "supports students self-

[2] Items O5, and S24 were excluded from the analysis due to a full response pattern in the more skilled teaching subgroup.

confidence.” This result suggests that student responses are triggered by the word “help,” (associated with “supports”) while less by the word “understand.”

Table 4.1

Resulting cumulative item ordering ranging from least complex teaching practices to most complex teaching practices

domain	item	teaching practice	*b*	$SE_{(b}$
SLC	O1	shows respect for students in behavior and language	−2.18	.353
SLC	O2	creates a relaxed atmosphere	−1.40	.266
SLC	S21	treats me with respect.	−1.15	.246
ECM	O7	ensures effective class management	−1.09	.242
SLC	O3	supports student self-confidence	−1.08	.242
SLC	S40	helps me if I do not understand.	−1.08	.242
ECM	S20	prepares his/her lesson well.	−1.05	.238
QOF	O9	explains the subject matter clearly	−1.02	.239
ECM	O5	ensures that the lesson runs smoothly	−.85	.225
SLC	O4	ensures mutual respect	−.77	.219
SLC	S8	ensures that I treat others with respect.	−.76	.219
QOF	O14	gives well-structured lessons	−.74	.219
SLC	S1	ensures that others treat me with respect.	−.68	.214
ECM	O8	uses learning time efficiently	−.64	.212
ECM	S6	answers my questions	−.51	.205
ATM	S23	ensures that I pay attention.	−.44	.202
ECM	S3	makes clear what I need to study for a test.	−.38	.198
ECM	S19	makes clear when I should have finished an assignment.	−.34	.197
QOF	S24	uses clear examples.	−.32	.195
QOF	S13	explains the purpose of the lesson.	−.31	.195
ECM	S39	involves me in the lesson.	−.30	.196
ECM	S26	applies clear rules.	−.25	.192
QOF	O6	checks during processing whether students are carrying out tasks properly	−.20	.193

--- continues next page ---

domain	item	teaching practice	*b*	*SE(b*
QOF	O15	clearly explains teaching tools and tasks	−.17	.193
QOF	O10	gives feedback to students	−.17	.190
ECM	S2	ensures that I use my time effectively.	−.13	.187
QOF	S33	ensures that I know the lesson goals.	−.13	.187
QOF	O11	involves all students in the lesson	−.07	.185
ATM	O13	encourages students to do their best	−.03	.184
ATM	S17	encourages me to think for myself.	.12	.178
ECM	S12	ensures that I keep working.	.23	.175
ATM	O19	asks questions that encourage students to think	.30	.172
ATM	O16	uses teaching methods that activate students	.37	.170
ATM	S30	stimulates my thinking.	.51	.168
ATM	O21	provides interactive instruction	.54	.167
QOF	O12	checks during instruction whether students have understood the subject matter	.57	.166
ATM	O20	has students think out loud	.60	.165
DII	S25	connects to what I am capable of.	.80	.161
DII	S34	checks whether I understood the subject matter.	.83	.161
TLS	O30	encourages students to apply what they have learned	1.00	.158
TLS	O31	encourages students to think critically	1.38	.155
TLS	S16	teaches me to check my own solutions.	1.43	.155
DII	S36	knows what I find difficult.	1.49	.154
DII	O23	checks whether the lesson objectives have been achieved	1.67	.155
TLS	O28	encourages the use of checking activities	1.82	.155
TLS	O29	teaches students to check solutions	1.87	.155
DII	O25	adapts processing of subject matter to student differences	2.30	.157
DII	O26	adapts instruction to relevant student differences	2.41	.158

Regarding the more advanced teaching practices the picture is still somewhat blurred. Because positive residual correlations tend to cluster around more complex teaching practices, the b-parameters of O25, O26, O28, O29, and 030 are biased. Inspection of two-by-two frequency tables of these items pairs gives the impression that O25 and O26 most plausibly are estimated as more complex than they actually are (i.e. the number in which they both score incorrect is higher than would be expected on the basis of the model), while O28, O29, and O30, are estimated as less complex as they actually are (i.e. the number in which these item pairs score correct is higher than would be expected on the basis of the model). In addition to this, the "My Teacher" questionnaire items describing more complex teaching practices more frequently show model violations and had to be removed. This is true in particular for the items within the domain teaching learning strategies. In this study only one questionnaire item within the domain teaching learning strategies remained: S16 "teaches me to check my own solutions." Currently, this complicates the comparison between the item parameters of the more complex teaching practices.

4.4.4 The correlation between methods

The correlation without item selection (*r* = .26) is lower than after removal of biased items (*r* = .34). The correlation without item selection is identical to the correlation recently reported by Maulana and Helms-Lorenz (2016). Thereby this sample reconfirmed their findings. Also, the results suggest that the low correlation is only partially dependent on items measuring different constructs. Most student items and classroom observation items are found to fit the one-dimensional ordering. Also, items measuring the same domains have similar complexity. When the few biased items are deleted the correlation increases with .08.

We have validated that the student questionnaire and classroom observation instrument measure one latent variable. Therefore, it is possible to average them into one composite evaluation score. These 'True' teaching skill estimates (θ_T) are given in Table 4.2. They are more reliable than those of either one method alone and they are directly related to the item complexities mentioned in Table 4.1. For instance, teachers receiving a theta score -.25 were observed to perform most items above item O6, and they would be advised to give attention to items O15, O10, S2, S33, and O11, because the complexity of these practices is close to the current performance level of teacher and class.

Table 4.2

Teachers evaluation scores. Theta (θ) values correspond to item-parameters in Table 4.1

raw score	raw student score[a]	raw observer score[a]	θ[a]	*SE*[b]	$n_{(\text{teachers})}$
20	11	9	-1.02	.411	1
21	16	5	-0.93	.412	1
22	15	6	-1.00	.402	1
25	12	13	-.80	.411	4
26	13	14	-.73	.411	1
27	9	18	-.72	.405	1
28	11	17	-.70	.408	2
29	16	13	-.55	.411	7
30	15	15	-.55	.403	4
31	14	17	-.47	.407	4
32	17	15	-.36	.412	5
33	19	14	-.25	.403	4
34	17	17	-.25	.414	5
35	18	17	-.18	.413	4
36	18	18	-.14	.410	10
37	19	18	.00	.414	11
38	19	19	.05	.406	11
39	19	20	.10	.413	8
40	18	22	.15	.417	13
41	19	22	.23	.408	15
42	20	23	.40	.420	7
43	20	23	.45	.414	11
44	18	26	.41	.428	1
45	21	24	.71	.416	5
46	20	26	.73	.431	3
47	21	26	1.01	.440	2

[a]. If multiple teachers had similar raw scores, the reported value is the mean.

[b]. If multiple teachers had similar raw scores, the reported value is the median.

4.5 Conclusion and Discussion

This study combined a student questionnaire (the “My Teacher” questionnaire) and observation instrument (the ICALT observation) and explored whether items of both instruments measure one latent variable, namely teaching skill. The general conclusion of this study is that students and observers agree on the complexity of similar teaching practices. This finding is inconsistent with previous speculations that student questionnaire and classroom observation methods must measure different constructs because they offer different perspectives. Clearly classroom observations and student questionnaires *may* result in different measurement. Questionnaires can address aspects of teaching which observers cannot readily observe (e.g., whether students understood the explanation). Also, classroom observation can evaluate aspects of teaching skill that students cannot reasonably evaluate (e.g., the quality of the lesson content and materials). But our results suggest that when observers and students evaluate aspects of teaching they both can observe, their ratings are psychometrically similar and one-dimensional.

This implies that the low correlation between classroom observation and student questionnaires cannot reasonably be explained as being due to both instruments measuring different constructs. While our results replicate the low correlation between students and observers, we also find that student questionnaire items and classroom observation items fit the same one-dimensional cumulative ordering. The correlation between evaluation methods is r = .26 and after removal of those few misfitting items increases only slightly to r = .34. This slight increase gives reason to doubt whether student questionnaires and classroom observation instruments when they would show perfect one-dimensional measurement—and measure exactly the same construct—would approach a correlation of r = 1.00. We need to consider other explanations as to why students and observers disagree on teachers’ teaching skill.

4.5.1 Alternative explanations of the low correlation between students and observers

What alternative explanations can be given for the unexpected low correlation? One important explanation may be the low reliability of one-time lesson visits, which provide an unrepresentative picture of the teachers’ teaching skill (Kane ,et al. 2012; Van der Lans, et al. 2016). On the other hand, Benton and Kashin (2012) and Marsh (2007) state that student questionnaires are more valid because students have observed all lessons and that they are with many. It might be unreasonable to expect that evaluation outcomes based on a one-time lesson visit by a single observer should be comparable with evaluation outcomes based on all lessons and observed by many observers. An example is Murray's (1983) study which correlated classroom observations of multiple lessons by multiple observers with student questionnaire data and reports a correlation of $r = .76$. This result suggests that the moderate correlation might be due to low reliability. This suggests that increasing the number of observers and the number of lessons observed could increase the correlation between classroom observation data and student questionnaire data.

4.5.2 Limitations

The study has some limitations which should be taken into account. One limitation involves the violations of local independence. Though these are few, they tend to concentrate around items within more complex domains. This is especially true for the observation instrument. Therefore, the teaching practices in the ICALT domains teaching learning strategies and differentiation in instruction are biased. It may be argued that due to accepting some violations of local independence the correlation is somewhat biased (either too optimistic or pessimistic). Another limitation concerns the cross-validation analysis which contained the same classroom observations twice and only varied the student ratings. It may be argued that the positive cross-validation result is due to using part of the data twice. Finally, a multilevel approach should be preferred when testing clustered data as students nested in classes. We note that the item parameters in Table 4.1 are estimated using multilevel techniques, but acknowledge that the Rasch model fit tests are not corrected for the clustered data structure. Methods to evaluate Rasch model assumptions within a multilevel framework are still in a developmental phase (e.g., de Boeck, et al. 2011). We consider our choice to randomly sample one student from each class as the best option currently available.

Chapter 5

How to decrease the risk of unreliable and invalid evaluations?

Abstract

Implementation of effective teacher evaluation procedures is a global challenge in which lowering the chances that teachers receive inaccurate evaluations is a pertinent goal. This study investigates the minimum number of observations required to guarantee that teachers receive feedback with modest reliability ($\boldsymbol{E}\rho^2 \geq .70$) and that any summative decisions about their professional career have high reliability ($\boldsymbol{E}\rho^2 \geq .90$). A sample of 198 classroom observations by 62 colleagues of 69 teachers working at eight schools reveals that reliable feedback requires at least 4 lesson visits by four different observers. Also results indicate that if only using classroom observation it is almost impossible to guarantee a reliability level sufficient for the use for summative decisions. The findings mirror those reported with other observation instruments. This study accordingly offers directions for how schools can implement classroom observation procedures cost-effectively.

Modified version of the article: Van der Lans, R. M., Van de Grift, W., J., C., M., Van Veen, K., & Fokkens-Bruinsma, M. (2016). Once is not enough: Establishing reliability criteria for feedback and evaluation decisions based on classroom observations. *Studies in Educational Evaluation, 50,* 88-95. doi: 10.1016/j.stueduc.2016.08.001

5.1 Introduction

The development and implementation of effective teacher evaluation is a global challenge, as various international policy documents and reports reveal (e.g., DfEE, 2012; Mourshed, Chijioke, & Barber, 2010; NCTQ, 2013). In all these policy documents teacher evaluation has a dual purpose: (1) identification and selection of ineffective teachers and (2) offering advice for improvement of teachers' teaching (Marzano, 2012). The global attention signals that many countries currently are interested in how to obtain more reliable information to support their summative decisions and formative feedback. That is, there is an interest in preventing wrong decisions about teacher selection and preventing the provision of wrong feedback about how to improve teaching effectiveness, since wrong decisions and feedback will harm individual teachers, and definitely not improve student learning outcomes.

Of these two purposes of teacher evaluation, the decisions about teacher selection currently receive most attention (e.g., Firestone, 2014; Winter & Cowen, 2014). Evidently, there is much at stake for individual teachers, who have worked hard to earn accreditation and succeed in classrooms. This gives researchers and policymakers the moral obligation to carefully consider the reliability of their decisions. Clearly, evaluations might be wrong and select teachers for dismissal which will prove to be effective. Also, evaluations might be wrong by not selecting teachers for dismissal which will prove ineffective. Currently, it is attempted to avoid wrongly removing effective teachers, but this automatically leads to a situation in which many ineffective teachers are wrongly retained (e.g., Winters & Cowen, 2013).

The provision of formative feedback has at first sight less severe personal consequences. Nevertheless, also feedback should be based on a representative picture of the teacher's true teaching skill. In general, educational policies rely on classroom observations specifically to target teachers who appear ineffective in some way and to provide them feedback (e.g., NCTQ, 2013). If these teachers show no improvement in their follow-ups, the policies suggest they should be selected for dismissal. Given these personal consequences, teachers deserve reliable feedback, such that it offers them a true opportunity to improve.

This study examines the reliability of classroom observation. Classroom observation is currently the most widely adopted teacher evaluation method (Strong, 2011). However, only few studies report on the reliability of these observation methods (e.g., Hill et al., 2012; Kane et al., 2012) and none of these studies relate reliability criteria to the two different purposes of teacher evaluation. This study seeks to determine if classroom observations can achieve a reasonable level of reliability to support both formative feedback and summative decisions, and if so, how many observations by how many separate observers are required.

5.2 Background

5.2.1 Evaluation reliability and purpose

An examination of validity and reliability should be related to the purpose for which the instruments will be used (Kane, 2006). In teacher evaluation, instruments generally are used for two different purposes. Therefore, different reliability criteria should apply to investigate whether instruments reliably support summative and formative evaluation decisions. However, studies examining classroom observation instruments rarely relate reliability criteria to the intended use of the instrument. For example, Hill et al. (2012) examine how much the reliability increases if evaluations incorporate multiple raters and lessons and seek "to achieve acceptable reliability" (p. 60), without clarifying what an acceptable level of reliability would be and whether that level might change if other evaluation purposes would apply. Similarly, Kane et al.'s (2012) influential report for the Measures of Effective Teaching (MET) project notes that:

> *"Not all decisions require high levels of reliability. Measures could be used many different ways: promotion decisions, retention decisions, compensation decisions, or low-stakes feedback intended to support improvement. Different uses necessitate different evidentiary standards and different levels of reliability (there is no uniform standard that applies to any envisioned use)."* (p. 13)

That is, though Kane et al. (2012) recognize that different evaluation purposes require different reliability criteria, they do not mention any specific criteria. In subsequent work for the MET project, Ho and Kane (2013) cite the reliability criterion $\boldsymbol{E}\rho^2 = .65$, without specifying the evaluation purpose for which this criterion would be appropriate. Because these studies do not set clear reliability criteria for different evaluation purposes, it appears that the reliability of classroom observations currently is determined by educational policies and the school principals' perceptions of what it takes to get a "reliable observation" for a given purpose.

To tie evaluation purposes to different reliability criteria, we adopt the criteria for both modest and high reliability formulated by Nunnally (1978). Therefore, we argue that modest reliability of $\boldsymbol{E}\rho^2 \geq .70$ suffices for formative feedback and other instances in which the stakes are relatively low. We suggest instead that a high reliability level of $\boldsymbol{E}\rho^2 \geq .90$ is the minimum criterion to use for summative decisions and instances in which “a great deal hinges on the exact score made by a person on a test” (Nunnally, 1978, p. 245).

5.2.2 Reliability of one-time lesson visits

Using multiple lesson visits is not standard practice in teacher evaluation, with some notable exceptions, such as the teacher advancement program (TAP) (Darling-Hammond, Amrein-Beardsley, Heartel, & Rothstein, 2012; Toch & Rothman, 2008). Yet it is common knowledge that one-time observations may be substantially biased by a bad moment or difficult class (e.g., Muijs, 2006; Shavelson & Dempsey-Atwood, 1976). In empirical studies of the reliability of a single lesson visit by a single observer, across different classroom observation instruments, the findings are fairly consistent. Ho and Kane (2013) report reliability coefficients between .27 and .45, depending on the type of observer (teacher peer or administrator); Kane et al. (2012) examine five classroom observation instruments and report coefficients of .37 or less. In Hill et al.’s (2012) study, the reliability coefficients for three different subscales of the Mathematical Quality of Instruction (MQI) hover between .37 and .46. That is, the reliability of single classroom observations is low and generally less than .50. Previous works suggest that at least three to four lesson visits are required to achieve even modest reliability ($\boldsymbol{E}\rho^2 \geq .70$) (Hill, et al. 2012; Ho & Kane, 2013; Kane, et al. 2012). Note that we use the notation $\boldsymbol{E}\rho^2$ to refer to the reliability coefficient. This notation is taken from Brennan (2001). The ρ^2 is the usual notation of reliability in classical test theory. The ***E*** signifies that the reported coefficient reflects the expected reliability. It is the reliability we would expect if the evaluation procedure is exactly repeated.

Beside low reliability, the validity of one time classroom visits has also been criticized on other grounds. One is that the person visiting also is the person judging and hence observation scores cannot be anonymous (Scriven, 1981). This makes the appointed evaluator most vulnerable to criticism (Popham, 1987; French-Lazovik, 1981) which in turn provides an incentive to give lenient scores (Centra, 1975; Weisberg, Sexton, Mulhern, & Keeling, 2009). Both Centra and Weisberg stated that an evaluation procedure which evaluates over 95% of the teachers as performing sufficient lacks validity. These studies show the necessity to clearly distinguish between those who observe and those who decide.

5.2.3 Potential evaluation procedures

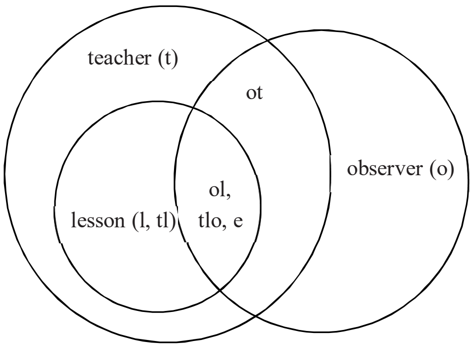
With the view that reliability is paramount to teacher evaluation and that single-lesson visits have unacceptably low levels of reliability, we discuss three evaluation procedures that might enhance the reliability of classroom observations, compare their pros and cons, and speculate whether their durable implementation in schools is realistic. The successful implementation of any evaluation procedure requires that it be cost effective and manageable for schools (Peterson, 2000). Ideally, an evaluation procedure would entail minimal organizational complexity but still provide sufficient guarantees that the resulting evaluations are reliable and fair. Furthermore, any implementation is restricted by the reality of the school organization. We consider three potential procedures: crossed, nested, and bias-confounded.

Crossed procedure. This complex evaluation procedure requires a group of observers to visit all lessons together. An example of the crossed procedure appears in Figure 5.1. At the left side of Figure 5.1 the evaluation procedure is visualized. Check boxes reflect that the observer visited the lesson.

Figure 5.1

A schematic representation of the crossed evaluation procedure (left) and the resulting variance decomposition (right)

	Observer		
	A	B	C
Teacher A			
lesson 1	✓	✓	✓
lesson 2	✓	✓	✓
lesson 3	✓	✓	✓
Teacher B			
lesson 4	✓	✓	✓
lesson 5	✓	✓	✓
lesson 6	✓	✓	✓
Teacher *n*			



At the right side, a Venn diagram representation is used to visualize the same procedure. In a Venn diagram each circle is a facet; areas where two circles overlap illustrates an interaction between two facets. The crossed procedure offers the most complete information, because it separates information about true differences across teachers (t) from any bias due to differences across lessons (l), bias due to observers (o), and bias due to their interaction (observer × teacher). In our notation, the "e" refers to "error." Furthermore, commas identify confounding facets. Confounds signal that variation is attributable to two or more facets such that the variation has no single interpretation. Hence the facet "lo, tlo, e" in Figure 5.1 reflects that this part of the variation in scores may be explained by lesson × observer interactions, by teacher × lesson × observer interactions, and by measurement error. As such this facet has no substantive interpretation.

This crossed evaluation procedure has been applied in previous studies of the reliability of classroom observations (Hill et al., 2012; Ho & Kane, 2013). It offers benefits, in that the crossed design offers information about the reliability of the evaluation, as well as details about the extent to which any particular bias affects reliability. If reliability is too low, the procedure reveals what to do: (1) add another observer, (2) prevent some particular observer from visiting some particular teacher, or (3) visit an additional lesson.

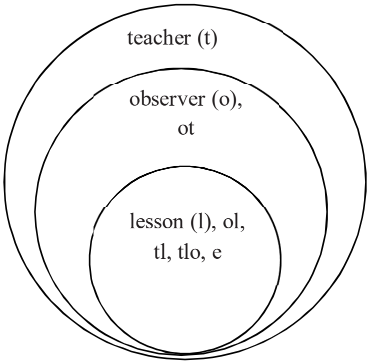
Despite its comprehensiveness, this evaluation procedure is unworkable in practice for most schools. In the hypothetical scenario where a school employs 50 teachers and requests three lesson visits with each teacher, it would demand 150 group visits by the same group of observers. The number of work hours also depends on the size of the group, but in this hypothetical case, if the group includes three observers, it would mean 450 hours of lesson observation. Most schools lack the financial resources to hire external observers, so the observation group likely consists of peer colleagues, team manager(s), or school principal(s). Each of these actors would have to perform 150 classroom observations, in addition to their existing obligations, and schedule these observations together. It is implausible that such procedures can be implemented successfully in schools, despite that this would be better from a psychometric point of view. In addition, likewise the one-time lesson visit procedure, also an appointed group of 'expert' observers will be vulnerable to criticism (French-Lazovik, 1981; Peterson, & Chenoweth, 1992). Because in the crossed procedure all teachers are evaluated by the same (small) group of observers and the observers will be more acquainted with some subjects, or befriend with some colleagues, it is likely that some of the teachers under evaluation will not feel that they are treated equally. Note also that in a research setting the strength of the crossed procedure is that it can take such observer-teacher interactions into account, but in the school practice there is no knowledge about such statistical models and currently it is unlikely that schools can take adequate actions to avoid tensions between colleagues when implementing the crossed procedure.

Nested procedure. As a more flexible approach (Figure 5.2), the nested procedure requires one group of observers to visit multiple lessons of one teacher together. The difference with the crossed procedure is that other teachers may be visited by other groups (see Figure 5.2). This flexibility comes with a price though. The procedure cannot reveal the extent to which reliability decreases due to observer × teacher interactions. Rather, the variance due to observer × teacher (ot) interactions sums with the variance due to observers (o), resulting in an "o, ot" facet that confounds two interpretations. That is, the variance in this facet might reflect differences among observers, or it could reflect differences in observer × teacher interactions.

Figure 5.2

A schematic representation of the nested evaluation procedure (left) and the resulting variance decomposition (right)

	Observer					
	A	B	C	D	E	F
Teacher A						
lesson 1	✓	✓	✓			
lesson 2	✓	✓	✓			
lesson 3	✓	✓	✓			
Teacher B						
lesson 4				✓	✓	✓
lesson 5				✓	✓	✓
lesson 6				✓	✓	✓
Teacher *n*						



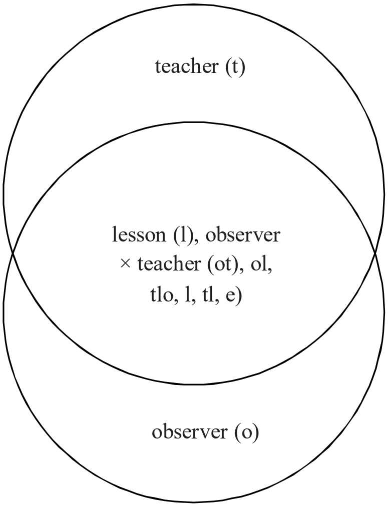
None of the research referred to in this study has used the nested procedure. It offers benefits in that it is more flexible with regard to who can perform the classroom observations in comparison to the crossed procedure. This flexibility is important since it provides the room to carefully select peer-observers for each teacher (French-Lazovik, 1981). Furthermore, it still provides some information about what to do if reliability is too low: add another observer or visit an additional lesson. However, the nested procedure is not any more efficient than the crossed procedure. Its implementation in our hypothetical, modest sized school would require different groups of observers to visit 150 lessons together, so if we again assume the groups include three peers, it still demands 450 hours of observation. Also, despite that now different groups may perform the classroom observations, schools still have to schedule group visits. They need to find groups willing to visit lessons together.

The bias-confounded crossed procedure. A yet lesser complex procedure involves the bias-confounded crossed procedure. In this procedure, teachers are grouped and teachers within the same group visit each other's lessons (Figure 5.3). The term "bias-confounded" signals that in this procedure all interaction facets are summed. The main difference with the nested and crossed procedures, is that a bias-confounded procedure allows for individual lesson visits. Thereby, the bias-confounded crossed procedure is much more efficient and requires only one-third of the lesson observations (i.e. 150 hours of observation) compared to the previously described crossed and nested procedures.

Figure 5.3

A schematic representation of the bias-confounded crossed procedure (left) and the resulting variance decomposition (right).

	Observer						
	A	B	C	D	E	F	G
Teacher A							
lesson 1		✓					
lesson 2			✓				
lesson 3				✓			
Teacher B							
lesson 4	✓						
lesson 5			✓				
lesson 6				✓			
…							
Teacher E							
lesson 7						✓	
lesson 8							✓
lesson *n*							
Teacher *n*							



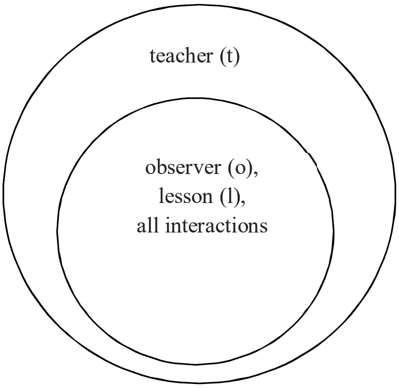
The increased efficiency comes with a cost. In comparison to the crossed procedure, if using this procedure, it is not possible to estimate the size of bias due to variations across lessons (facet l). The facet "l" is summed with the facet observer × teacher interaction (ot) and all its interactions "ol", "tl" and "tlo" and further confounds with measurement error. In comparison to nested procedure, the crossed procedures are less flexible. The bias-confounded crossed procedure is more rigid, because it requires that teachers are grouped together and teachers within one group cannot simply be scheduled to visit lessons of any random colleague but only those of the few colleagues within his/her group. This restriction limits easy implementation in schools. However, it does provide additional information about possible sources of bias. As shown in the Venn diagram in Figure 5.3, if using a bias-confounded crossed procedure, it is possible to estimate bias due to teacher × observer interactions (facet ot) and to separate this bias from the bias due to observers (o). This is not possible in the bias-confounded nested procedure discussed next.

Bias-confounded nested procedure. The least complex procedure, or what we refer to as the bias-confounded nested procedure, has multiple observers visit one teacher's classrooms individually (see Figure 5.3). This procedure cannot indicate why classroom observations might emerge as unreliable. Rather, differences across lessons sum with differences among observers resulting in the "o, l, lo, to, tl, tlo, e". That is, all variance not attributable to differences in teaching are represented in a single error facet.

Figure 5.4

A schematic representation of the bias-confounded nested procedure (left) and the resulting variance decomposition (right).

	Observer					
	A	B	C	D	E	F
Teacher 1						
lesson 1	✓					
lesson 2		✓				
lesson 3			✓			
Teacher 2						
lesson 4				✓		
lesson 5					✓	
lesson 6						✓
Teacher *n*						



This procedure was examined by Kane et al. (2012) and advocated by Ho and Kane (2013, table 10). Its greatest benefit is its flexibility (anyone who receives training can perform a visit) in combination with an increased efficiency (requires fewer visits). Its greatest disadvantage is that the procedure provides no information about what specific actions can be taken in case where reliability is found too low. In our hypothetical example, with three peers visiting three lessons, the procedure requires just 150 hours of observation instead of the 450 hours required by the previous two procedures. Also, schools do not have to find groups willing to visit multiple lesson taught by the same teacher together. Still, even this evaluation procedure demands considerable commitment from the school.

In summary, the crossed evaluation procedure, in which observers visit all the lessons together as a group (optimal situation from a psychometric perspective), is unrealistic for schools. Successful implementation instead requires a reduction of organizational complexity. The resulting situation is less than optimal, but more realistic, and it suffices to estimate the reliability of classroom observations. In this study, we have implemented the bias-confounded crossed procedure for this study.

5.3 Study aims and research questions

We explore the potential reliability of an evaluation design, as it has been implemented by actual schools. In so doing, we seek to replicate previous findings by Kane et al. (2012), Hill et al. (2012), and Ho and Kane (2013) that suggest that incorporating multiple lesson visits by multiple observers substantially increases reliability. This study also expands those previous works, by estimating the gains in reliability relative to certain absolute cutoffs (i.e., modest reliability $\boldsymbol{E}\rho^2 = .70$ and high reliability $\boldsymbol{E}\rho^2 = .90$) and explicitly relating the criteria to the different purposes of an evaluation, namely, formative feedback and summative decision, respectively.

Our focal research questions are as follows:

1. How many classroom observations by peers are required to achieve modest reliability and support formative feedback?
2. How many classroom observations by peers are required to achieve high reliability and support summative decisions?

5.3 Method

To investigate the research questions, peer observers in eight different schools across the Netherlands received training to perform observations of their colleagues. This type of collegial visitation fits the purpose of formative feedback, as well as current policies in the Netherlands (OCW, 2013a). The participating teachers each received three lesson visits by three different peers, after which we computed an evaluation score that could range from 0 to 31, such that 0 indicates the teacher poorly performed all of the teaching practices listed in the instrument, and 31 indicated the teacher competently performed all of these practices. On the basis of this score, the teachers received feedback in a 20-minute, face-to-face conversation with the researcher, focused on their current teaching skills and the most likely options for improving their teaching.

5.3.1 Sample

Three different peers each observed a lesson taught by each teacher. The peers ensured that their lesson visits were scheduled for the same class. Using this procedure, we obtained 198 lesson observations of 69 teachers by 62 peers working at eight different schools across the Netherlands. The number of lesson observations is smaller than three times the number of teachers due to situational circumstances, such as when one of the three peers or the specific teacher was temporarily unavailable to perform or have lesson visits. Thus, 14 teachers were observed on only two occasions.

Teachers. Teacher experience ranged from 1 to 40 years (M = 13 years, SD = 10 years), and 62.1% of them were men. The non-representative gender distribution prompted us to check if male teachers might be evaluated differently than their female counterparts. An analysis of variance (ANOVA) revealed a negligible difference between male and female teachers ($F(1, 196) = 1.756$, $p = .18$). In addition, the teachers engaged in all available educational types: preparatory secondary vocational education (20.7%), senior general secondary education (46.5%), and university preparatory education (26.3%). The observed subjects were math (22%), history (21%), Dutch (20%), English (20%), and geography (4%), as well as German, Latin, economy, social sciences, science, religion, and construction (all ≤ 2%). Classroom observations took place between March and June 2014 and between February and June 2015.

Peer observers. Observers' teaching experience ranged from 1 to 40 years (M = 18 years, SD = 11 years), and 71.7% of them were males. Again, we checked if the unequal division of male and female teachers affected the overall evaluation results; the one-way ANOVA suggested no difference between male and female observers ($F(1, 196) = .01$, $p = .97$) or any indications of observer-gender × teacher-gender interactions ($F(1, 194) = .69$, $p = .56$). So, it seems likely that similar evaluation scores will be obtained in case that the division between males and females is more equal. In most instances, the peer observers were full-time teachers, though not all of them taught full-time. In modern Dutch schools, team managers frequently are part-time teachers, such that the boundaries between peer-teacher and peer-manager are permeable. We use the word "peer" to refer to school personnel, all of whom have (previous) teaching experience.

5.3.2 Instrument

The International Comparative Analysis of Learning and Teaching is a Rasch-scaled observation instrument (Van de Grift, Helms-Lorenz & Maulana, 2014; Van der Lans, Van de Grift, & Van Veen, 2016). The most recent update of the instrument includes 31 items, each representing an effective teaching practice, such as "uses teaching methods that activate students." The items span six domains: safe learning climate, classroom management, clear instruction, activating students, teaching learning strategies, and differentiation (for details, see Van de Grift, 2014). Observers rated the items as either 0 = "insufficient" or 1 = "sufficient."

5.3.3 Procedures and training

The research procedure sought to simulate what a real implementation in schools would involve. That is, schools have limited time and resources for observation training, so for this study, the training lasted four hours, and observers were considered "limitedly trained." All colleague-teachers could participate in the training irrespective of their previous experiences with classroom observation. Also, we did not apply any tests or certification systems to prevent peer observers with insufficient inter-rater reliability from entering the classrooms; any peer who participated in the training was accepted as an observer, irrespective of his or her performance. These decisions are made because most schools have limited or no access to statistics, such that a real implementation would not involve the computation of inter-rater reliabilities. Also, schools are social organizations with their own group dynamics (Peterson, 2000). It is unlikely that they will (and can afford to) exclude willing peers from observing lessons. Therefore, this research aims to achieve sufficient reliability, given that schools decide to have all willing teachers participate in collegial visitation.

Observation training. The observation training involved a half-hour introduction to the instrument, after which the observers scored two lesson videos, each 20 minutes in length. Four different videos were available for the training, two in each training session. The videos were not randomly assigned; rather, in spring 2014, we used videos 1 and 2, and in spring 2015, we used videos 3 and 4. In both years, the training started with an easy video followed by one that was more difficult to score. After each video, we calculated the percentages of observer agreement and discussed any problematic or confusing items. The videos of similar difficulty levels achieved similar consensus percentages: video 1 (74%) versus video 3 (75%) and video 2 (65%) versus video 4 (66%). Depending on the group, we also provided time to allow the trainees to express any insecurities about observing their peers.

5.3.4 Data preparation

During their observations, the peer observers were instructed to score as many items as possible. If a teaching practice was not observed, they had to decide whether in that lesson situation, the teacher should have used the practice, in which case the item was scored insufficient, or if the lesson situation did not allow for its performance, in which case the observers would leave the item blank. Of all item responses, only 3% were reported missing, so we considered them missing at random. We used procedures outlined by Raju, et al. (2006) to estimate an internal consistency coefficient similar to Cronbach's alpha. The internal consistency was high, $\rho_{(xx')} = .90$. However, consistency at the higher end of the measurement scale was considerably lower. Specifically, for raw scores of 30 and 31, the coefficient was less than $\rho_{(xx')} = .70$, so the evaluations did not consistently discriminate between the most excellent teachers.

5.3.5 Analysis

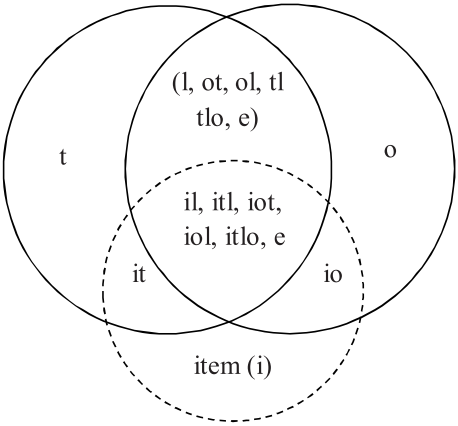
To examine the effect of adding additional peer observers, we used a Generalizability in Item Response Model (GIRT) methodology, as described by Briggs and Wilson (2007) and Choi (2013). The study design involves lessons (l) nested in observers (o) and teachers (t), crossed with items (i) (abbreviated (l:(o × t)) × i). The Venn diagram in Figure 5.5 is identical to the bias-confounded crossed procedure in Figure 5.3, except that it adds the item (i) facet, to describe the difference in chance on a positive score on the item describing the least complex teaching practice and the most complex teaching practice. This item facet is not a form of bias, because it describes a rank ordering in items identical for all teachers. In contrast, the facets item × observer (io) and item × teacher (it) should be interpreted as biases; they describe the degree to which the rank ordering is not identical for all teachers. For convenience, we refer to the facet of observer × teacher (ot), though more accurately, this facet is the sum of variation due to lessons (l), the observer × teacher interaction, and all interactions of facet lesson with the other facets.

To estimate the reliability coefficients, we used a two-step procedure. First, the generalizability (g-) study examines the amount of variation for which each facet, “t,” “o,” “i,”, and their interactions can account. Second, the decision (d-) study examines the increase in reliability expected from adding more levels to a facet (Brennan, 2001). The Appendix F (Technical appendix) provides a more detailed explanation of the GIRT analysis.

G-study. The facets “o,” “t,” and “i" and their interactions were estimated using a multi-facet Rasch (1960) model, with the R package lme4 (Bates, et al., 2014). This package is a general statistical software package. Descriptions of how to formulate and estimate Rasch models using lme4 are available in de Boeck et al. (2011).

Figure 5.5

Venn diagram of the implemented bias-confounded crossed procedure with the item facet



D-study. The d-study examines the increase in reliability achieved by adding more peer observers. We studied two cut-off points, $\boldsymbol{E}\rho^2 = .70$ and $\boldsymbol{E}\rho^2 = .90$, and estimated how many observers would be required to achieve these levels. The logic underlying the d-study is that if the variance due to observers is large (e.g., 50% of total variance) and the number of observers is small, any particular observer adds considerably to the shifts across evaluation scores. Consequently, the relative weight of the observer facet (i.e., bias) should be greater, and the average evaluation score is unreliable. However, if the variance due to observers remains similar, even with an increasing number of observers, the average evaluation score depends less on any particular observer. The relative weight of the observer facet then decreases, and the average evaluation score becomes more reliable. To estimate the relative increase in reliability with additional observers, a d-study assumes that the observer variance determined from the g-study is a true, unchanging reality, covering the complete range (or universe) of disagreement across classroom observers (Brennan, 2001). That is, this variance percentage can be expected with any number of observers. The d-study then varies the number of observers ($n_{(o)}$), thereby changing the relative weight of the observer facet in the reliability equation, to estimate the reliability levels with more or fewer observers. The d-study design is o × t × I. The capitalized "I" signifies that we consider the facet "items" as fixed, consistent with item response theory (Briggs & Wilson, 2007).

5.4 Results

To address the research questions, regarding how many classroom observations by limitedly trained peers are required to provide teachers with sufficiently reliable evaluations for the purposes of formative feedback ($\boldsymbol{E}\rho^2 \geq .70$) or summative decisions ($\boldsymbol{E}\rho^2 \geq .90$), we summarize the results of the G-study, with the design o × t × i, in Table 5.1.

Table 5.1

Variance Decomposition for the Multifacet Rasch Model

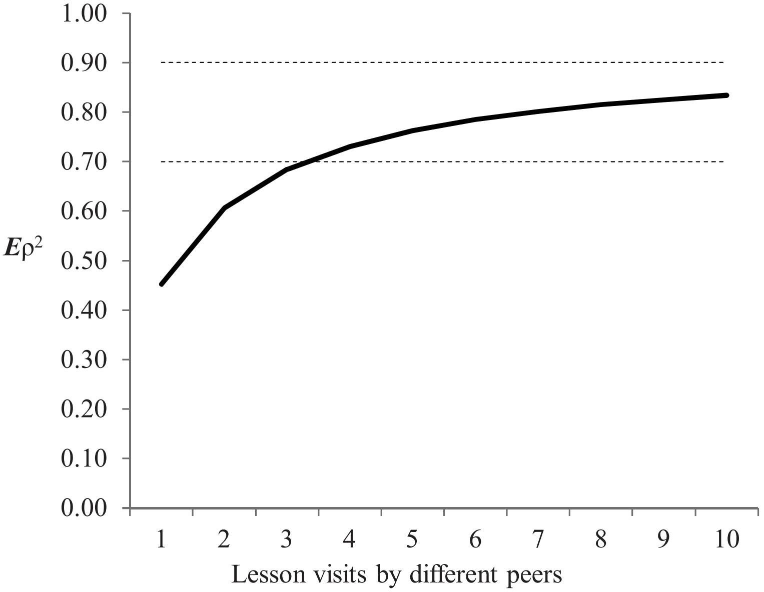
	$\boldsymbol{E}(\sigma^2)$	%
teacher (t)	1.29	0.22
observer (o)	0.37	0.06
item (i)	2.03	0.35
observer × teacher (l:ot)	1.07	0.19
item × teacher (it)	0.30	0.05
item × observer (io)	0.70	0.12
item × observer × teacher, e (i(ot:l), e)	.00	.00

Note. The number of estimated facets is different from the number published in Studies on Educational Evaluation. After the article was published we discovered that we had needlessly confounded two facets (observer and observer × teacher), which in fact could be separated.

As these results reveal, 22% of the variation in observed scores is due to true differences in teachers' skill. Furthermore, evaluations of the same teacher can vary substantially among observers: i.e. sum of facets o and ot. The variation due to observers is as great as the variation due to true differences in teaching skill. This substantial variation between observations—which, in the bias-confounded procedure, reflects the combined variance due to observers and lessons—is consistent with previous results (Hill et al., 2012; Ho & Kane, 2012; Kane et al., 2012). However, our results diverge in one important respect from previous findings: By using the GIRT method, we include the item (i) facet. This GIRT-based method includes more information for estimating evaluation scores than previous estimation techniques have, which should improve the reliability of the evaluation scores.

Figure 5.6

Expected increase in reliability (***E***ρ^2) with increasing numbers of lesson visits by different peer observers.



Note. These estimates differ slightly from the results published in Studies on Educational Evaluation. They differ in terms of height, not in terms of direction. Resolving the confound (see note Table 5.1) also improved and changed the estimation of the teacher facet (t) and this decreased the estimated reliability coefficients by approximately .06.

This improvement is reflected in the expected reliability of an evaluation based on a single lesson visit, which is slightly higher than in previous works, yet still only $\boldsymbol{E}\rho^2$ = .45 (Figure 5.6). Figure 5.6 depicts how much this reliability is expected to increase with additional peer observers. To exceed the modest reliability criterion for formative feedback: i.e., reliability ≤ .70, a minimum of four lesson visits is required ($\boldsymbol{E}\rho^2$ = .72). The number of lesson visits required to exceed the high reliability criterion for summative decisions ($\boldsymbol{E}\rho^2$ = .90) exceeds 20 and it seems though not possible to reach this criterion if using only classroom observations. After 10 lesson observations ($\boldsymbol{E}\rho^2$ = .83) the relative increase in reliability of additional observations becomes negligibly small (i.e. <.01).

5.5 Conclusion and Discussion

This study investigates whether increasing the number of lesson visits and the number of peer observers also increases the reliability of teacher evaluation. Our findings indicate that reliable formative feedback demands observations of at least 4 different lessons by different peers, and reliable summative decisions demand that evaluators gather more than only classroom observations. These results align with previous findings that predict modest reliability when four different observers visit one another's lessons (e.g., Hill et al., 2012; Ho & Kane, 2013). This study further shows that this reliability also can be achieved with less complex evaluation procedures and without overly restrictive training protocols. After approximately 10 visits, additional classroom observations add almost negligible amounts of new information (increase in reliability less than .01). Hence, these values, of at least 4 and more than 10 visits, therefore are highly relevant for real-world evaluation practices by schools. They provide preliminary insights for how to start implementing classroom observations using cost-effective, manageable procedures, while still ensuring generally acceptable reliability.

The findings share similarities with results presented about five other classroom observation instruments 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) (see Table 5.2).

Table 5.2

Reliability indices reported for the ICALT in comparison with reliability indices reported for the FFT, CLASS, UTOP, MQI and PLATO (Kane, et al 2012, Table 11)

	one visit	Two visits	three visits	Four visits
ICALT	.45	.60	.68	.72
FFT	.37	.53	-	.67
CLASS	.31	.47	-	.63
UTOP	.30	.46	-	.63
MQI	.14	.24	-	.34
PLATO	.34	.50	-	.67

Therefore, the values of ≥ 4 (modest reliability) does not appear unique to the observation instrument that we applied; rather, it seems to be broadly characteristic of classroom observation instruments in general.

5.5.1 Alternative procedures to increase reliability

The number of lesson visits required to establish an acceptable reliability for summative evaluation is estimated as considerably more than 10 visits with each teacher. This currently seems an impossibly great number to achieve for schools and brings us to the question, What alternatives exist to increase the reliability of teacher evaluations? We discuss some possible directions, which should be subject to further research.

Kane et al. (2012) report that evaluations that combine different measures (e.g., student ratings, classroom observations, student achievement) are more reliable than evaluations based on classroom observations only. Such combinations accordingly might reduce the number of observers required. Alternatively, further development and improvement of the instrument we used could reduce these thresholds too. Our results suggest that classroom observations are currently biased by an item × teacher interaction and item × observer interaction (together 17% of the total variation). If this facet could be reduced to approximately 0%, the reliability will slightly increase but by no more than approximately .01. Finally, most previous studies in this field rely on procedures involving videotaped lessons (Hill et al., 2012; Ho & Kane, 2013; Kane et al., 2012). Videotaping technologies suggest some great potential for increasing flexibility, because the videos of teachers could be watched by observers at any time, so the observation hours could be scheduled more flexibly. However, they also require schools to possess appropriate technical skills and equipment, particularly to ensure clear recordings of teachers' speech. The use of video also raises questions about whether these evaluations would be identical to evaluations based on actual lesson visits.

5.5.2 Limitations

This study has several limitations. First, the evaluation procedure (study design) did not incorporate differences across classes. A teacher's performance plausibly fluctuates from class to class, and the justification of summative decisions demands evidence of systematically poor or excellent performance across multiple classes, so the by Figure 5.6 estimated numbers of required lesson visits to achieve certain levels of reliability still are probably too low. Second, the current analysis estimates the increase in observation reliability for teachers with "average" teaching skill, to establish a single value of required visits. For performance at the extremes, generalizability theory instead predicts the need for fewer required observations (Brennan, 2001). Third, the terms modest and high reliability remain highly subjective. Although we use statistical cutoffs to define them, those very thresholds need to be subject to scrutiny and debate. Our criteria for reliability, following Nunnally (1978), have achieved wide acceptance. However, even Nunnally describes his criterion of .90 as a minimum to be tolerated and suggests that .95 should be the standard. Such a standard obviously would generate an even higher number of required lesson visits

Chapter 6

Individual differences in teacher development and their consequences for teacher evaluation

Abstract

Recent research provides some evidence indicating that teachers' development of effective teaching practices can be described using a six stage model. However, crucial to stage theories is its cumulative ordering to be reasonably valid for any individual teacher. Using teaching observation, this study explores whether and how many lessons show substantial deviations from the predicted cumulative ordering. Furthermore, we examine whether deviations cluster around some specific teachers. Three lessons of each teacher were observed each by another observer. The sample consists of 198 lessons taught by 69 teachers and observed by 67 observers. Rasch analysis again confirmed the general stage ordering. Using person fit statistics, 15% of the lessons are identified as showing substantial deviations from the cumulative ordering. However, no teacher consistently shows deviations, suggesting that these deviations are incidental. Implications for the evaluation of teachers' teaching skill are discussed.

Based on the manuscript: Van der Lans, R. M., Van de Grift, W., J., C., M., & Van Veen, K. (2016). *An examination of individual differences in teachers' teaching behavior development and their implications for teacher evaluation.* Manuscript submitted for publication.

6.1 Introduction

Teacher observation has become the crucial component in current policy efforts to improve education (Strong, 2011). This widespread acceptance of teacher observation among evaluators is partly due to its promise that it can provide teachers feedback regarding their development and improvement of effective teaching. If teacher observation is valued for its potential to provide formative feedback, then it seems logical to establish connections between teacher observation instruments and theory about teacher development. Several theories of teacher development have been proposed (e.g., Berliner, 2001; Fuller, 1969). Some studies also discussed the development of effective teaching connecting literature on teacher effectiveness with theory on teacher development (Antoniou, Kyriakides, & Creemers, 2015). This study builds on both of these works and hypothesizes that teachers' development of effective teaching broadly follows six stages. Some empirical evidence indicates that this six stage theory may provide a valid description of the development of most teachers (Van de Grift, Helms-Lorenz & Maulana, 2014; Van der Lans, Van de Grift & Van Veen, 2016). Thus, most – but maybe not all teachers – can be described using these six stages.

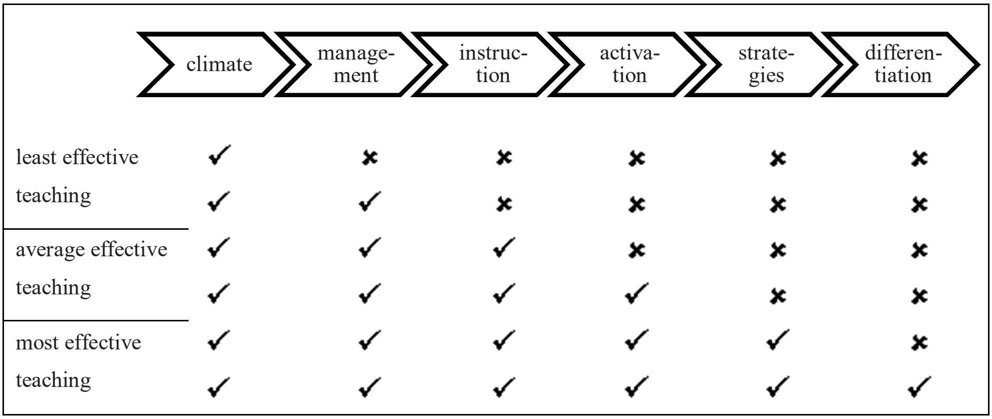
This study explores whether and how many teachers' development does not align with these six stages (and thus form so-called 'exceptions'). Some have speculated that there are individual differences in teacher development (e.g., Sternberg & Horvath, 1995). From the perspective of teacher evaluation the possibility and degree to which individual teachers deviate from the predicted stage ordering needs examination, because if substantial individual differences exist, the current six stage model will give teachers ill-informed advice regarding development and improvement of their teaching.

6.2 Background

This study builds on previous work: e.g., Van de Grift, Helms-Lorenz & Maualana (2014) and Van der Lans, Van de Grift & Van Veen (2016) which main premise is that teachers' development of effective teaching practice follows six broad stages. The hypothesized six stages predict that teachers (1) begin with learning how to establish a safe learning climate, (2) then proceed with learning how to efficiently manage a classroom, (3) develop skills in instruction, (4) then develop skills in more advanced teaching methods, including methods to activate students, (5) proceed with learning skills about how to teach students learning strategies, and finally (6) develop skill in differentiation of instruction. The proposed stages are presented in Figure 6.1.

Figure 6.1

Stage-wise progression in teacher development of effective teaching.



	climate	manage-ment	instruc-tion	activa-tion	strate-gies	differen-tiation
least effective teaching	✓	✗	✗	✗	✗	✗
	✓	✓	✗	✗	✗	✗
average effective teaching	✓	✓	✓	✗	✗	✗
	✓	✓	✓	✓	✗	✗
most effective teaching	✓	✓	✓	✓	✓	✗
	✓	✓	✓	✓	✓	✓

This six stage model serves as a heuristic model. The claims based on the model are modest. They include that (1) effective teaching practices can be ordered cumulatively from less complex to more complex and (2) that broadly this ordering reflects the six stages. So, we will still consider the model to give an inappropriate description if: (A) some teaching practices typical to a more complex stage develop before some of the practices typical to a less complex stage, (B) some specific teachers show a substantially different developmental ordering. The model in Figure 6.1 is a simplification of a more comprehensive developmental process.

The hypothesized six stages may be interpreted as a more detailed description of Francis Fuller's (1969) three stage theory of teacher development. Fuller proposed that beginning teachers (1) first develop skills how to establish relationships with students, (2) then start rethinking and adapting their classroom management and instruction methods, and (3) finally rethink their teaching methods and how they may increase their impact on student learning. Fuller developed her theory in the context of student teachers and beginning teachers and in particular Fuller's final stage may be perceived as lacking specificity. This makes her theory less valuable for evaluation of more experienced teachers – whom may be expected to show skill in Fuller's first two stages. The six stage theory addressed in this study, therefore, further elaborates Fuller's third stage.

Empirical work on the basis of lesson observations by external observers (Van de Grift, Helms-Lorenz & Maulana, 2014) and on the basis of students' ratings of teaching practice (Maulana, Helms-Lorenz, & Van de Grift, 2015) have both provided evidence in support of this cumulative six stage ordering in teaching practices. In addition, Kyriakides, Creemers, and Antaniou, (2009) report about a similar cumulative ordering using a student questionnaire related to five factors included in their dynamic model. Furthermore, the predicted cumulative ordering also fits with some auxiliary assumptions behind other theories and models currently in use to evaluate teachers. For example, the theory behind the Classroom Assessment Scoring System (CLASS) (e.g., Hamre et al., 2013) argues that teacher-student relationships are the key elements and building blocks of learning and teaching. Using Bowlby's attachment theory, they posit that in classrooms where students feel safe the students will become more self-reliant and confident to take risks, explore, and learn. This auxiliary assumption neatly fits with our six stages in which "safe learning climate" is the foundation of all further development in teaching effectiveness. As another example of such auxiliary assumptions, theory that describes teachers' development in terms of teaching expertise (e.g., Berliner, 2001; Sternberg & Horvath, 1995) assumes that teachers need to develop routines in more basic teaching practices – and most particularly classroom management and instruction – before they can start developing more advanced teaching methods.

6.2.1 Complexities in the evaluation of stage models

When investigating individual differences in development, it is important to distinguish between variation in performance and variation in development. Literature provides ample evidence that teachers' proficiency varies between classrooms (e.g., Goldhaber & Hansen, 2013). For the particular case of teacher observation methods, it is well documented that teachers' performance varies substantially between any two lessons (e.g., Kane, et al. 2012). Thus, the same teacher may not succeed to use teaching practices beyond the stage "quality of instruction" when teaching one class of students, while succeeding the implement teaching practices in the stage "teaching learning strategies" when teaching another class. In this sense, teaching is not completely dependent on the teacher, but also on the class (see for example; Doyle, 1983, 2009; Kennedy, 2010). This dependency on the class creates fluctuations as one teacher may move backwards and forwards within the model. In addition to the class, according to Borko (2004) and Dall'Alba and Sandberg

(2006) the context is also important. Performance varies also due to specific circumstances within the class. For example, Shulman's (1987) concept of pedagogical content knowledge predicts that teachers are more competent teaching certain subject matter content than other subject matter content. Thus, the same teacher teaching the same class of students may not succeed to use teaching practices beyond the stage "quality of instruction" when teaching particular subject matter content, yet succeeds to implement teaching practices at the stage "teaching learning strategies" when teaching other subject matter content. The Figure 6.2 below illustrates the distinction between performance and development. In this Figure Teacher A shows considerable variation in performance, while not showing any deviation from the cumulative ordering; Teacher B shows similar performance, while showing considerable deviations from the predicted cumulative ordering.

Figure 6.2

The distinction between variation in performance and variation in development. Teacher A varies in performance but shows no deviation from the cumulative developmental ordering. Teacher B shows no variation in performance, but deviates from the cumulative developmental ordering.

	climate	management	instruction	activation	strategies	differentiation
Teacher A	✓	✓	✗	✗	✗	✗
Teacher A	✓	✓	✓	✓	✓	✓
Teacher A	✓	✓	✓	✓	✗	✗
Teacher B	✗	✗	✓	✓	✗	✗
Teacher B	✗	✗	✗	✓	✓	✗
Teacher B	✗	✗	✓	✓	✗	✗

Our argument here is that this variation in performance does not necessarily reflect differences in developmental pathways. Theories about development generally assume that some elements of the profession are less complex and prerequisite to develop other more complex elements of the profession. The linear progression proposed by stage theories, therefore, is only falsified when it can be demonstrated that performance at 'higher' stages is possible when performance in 'lower' stages has not occurred. We will speak about ‘exceptional lessons’ and ‘deviations’ when teachers use more complex practices without also implementing less complex teaching practices.

6.2.2 One size fits all?

Critical of whether stage-theories are applicable is their degree of rigidity. Some theories of teacher development have completely abandoned the idea of stages (e.g., Berliner, 2001; Dall'Alba & Sandberg, 2006; Day et al., 2007; Huberman, 1993; Sternberg & Horvath, 1995), because they are unconvinced that the development of all teachers follow an identical sequence from less to more complex. Evidently it seems unrealistic to think that all teachers’ development follows completely identical steps and that each step needs to be completely completed before turning to the next step. In this study, we will turn to the question whether all teachers follow this same cumulative ordering.

Taken together this study addresses the following two research questions:

1. How many observed teaching situations show substantial deviations from the six stage ordering?
2. Do deviating teaching situations cluster with some particular teachers?

6.3 Method

6.3.1 Sample and procedures

The sample counted 198 lesson observations of 69 teachers by 62 observers. The study design grouped four teachers teaching the same class. In each group all teachers visited one lesson of their colleague teachers. Note that this design should have resulted in 69 times 4 is 207 lesson observations. Some teachers received only two lesson visits because of situational circumstances inside the schools. In total 54 teachers (78%) received three lessons visits by three different colleagues, another 14 teachers (20%) received two lesson visits by two colleagues and two teachers received four lesson visits by different colleagues. This study design should also lead to an equal number of observers and teachers (which is not completely the case). Schools substantially varied in how they assigned observers. One school decided not to use colleague teachers at all. In this particular school, all 36 lesson visits of 12 teachers were performed by the same three teacher coaches. Regarding the observers, 18 colleagues (29%) visited one lesson, another 7 colleagues (11%) visited two lessons, 24 colleagues (39%) visited three lessons, and 14 colleagues (22%) visited more than three lessons.

Teacher experience ranged from 1 to 40 years (*M* = 13 years and *SD* = 10 years) and 62.1% of the teachers were male. An ANOVA test confirmed that the unrepresentative amount of male teachers has few implications: the difference between males and females is negligible (*F*(1, 196) = 1.756, *p* = .18). The teacher observations took place between March and June 2014 and between February and June 2015. All observers also had teaching experience ranging from 1 to 40 years (*M* = 18 years and *SD* = 11 years) and again 71.7% of the observers were male. A one-way ANOVA test suggested neither difference between male and female observers (*F*(1, 196) = .01, *p* = .97), nor indications of observer-gender × teacher-gender interactions (*F*(1, 194) = .69, *p* = .56.

6.3.2 Instrument and observation training

The International Comparative Analysis of Learning and Teaching (ICALT) is a Rasch scaled observation instrument (Van de Grift, Helms-Lorenz & Maulana, 2014). In its most recent update the instrument includes 31 items each representing an effective teaching practice, such as *"uses teaching methods that activate students."* The items refer to six domains or stages: safe learning climate, efficient classroom management, quality of instruction, activating teaching methods, teaching learning strategies, and differentiation. Observers rated the items as 0 = "insufficient," and 1 = "sufficient").

The observation training involved half an hour introduction to the instrument after which two lesson videos of each 20 minutes were scored. Four different videos were used for training, two in each training session. The videos were not randomly assigned to observer groups. In spring 2014 we applied video 1 and 2, in spring 2015 video 3 and 4. In both years, the training started with an easy video followed by one video more difficult to score. Videos of similar difficulty achieved similar consensus percentages: video 1 (74%) versus video 3 (75%) and video 2 (65%) versus video 4 (66%). After each video percentages of observer agreement were computed and problematic or confusing items were discussed and clarified.

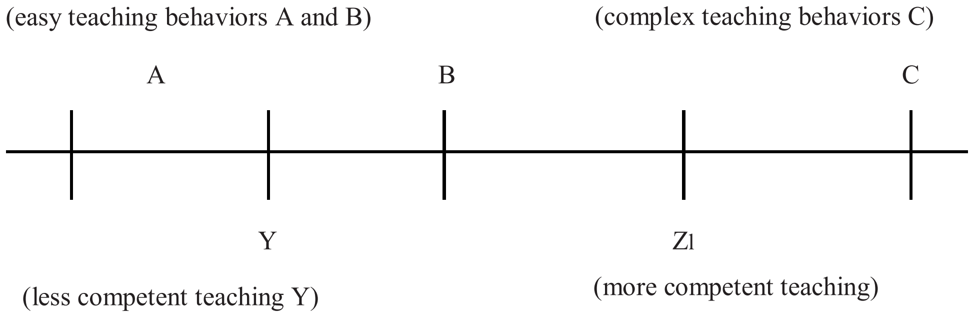
During the training observers are instructed and stimulated to score as many items as possible. If teaching practices are not observed, observers have to decide whether there were situations where the teacher should have used the practice – in which instance the item was scored insufficient – or whether there were no such situations – in which instance observers are instructed to leave the item blank. Of all item responses, 3% were reported missing. We considered these missing values missing at random (MAR).

6.3.3 Data analysis plan

To examine deviations from the cumulative ordering it is required to establish a baseline – i.e., the cumulative ordering within this particular sample – and verify whether it fits with the six stages. The Rasch model, which is specifically developed to estimate cumulative ordering among items (Bond & Fox, 2007), is applied to provide this baseline cumulative ordering. The software R-package eRm (Mair & Hatzinger, 2007) is used to estimate the baseline cumulative ordering in teaching practices for this particular sample. Next some statistics that evaluate whether particular teachers substantially deviate from this baseline cumulative ordering are required. For this second step two person fit statistics are applied: GNORMED (Meijer, 1994) and the χ^2-test of person fit available in eRm (Mair & Hatzinger, 2007). GNORMED is a nonparametric statistic based on the number of Guttman errors. To understand Guttman errors, it is required to explain that a (cumulative) measurement scale simultaneously describes two types of dominance relationships (Van Schuur, 2011). First it is possible that the teacher dominates the item – i.e. the teacher has more skill in teaching then is required to perform the teaching practice described by the item. Second, it is possible that the item dominates the teacher – i.e. the teaching practice described by the item requires more skill then the teacher has (Figure 6.3).

Figure 6.3

An illustration of the cumulative one-dimensional scale in which teacher Y dominates the teaching practice A, but is dominated by the teaching practices B and C.


(easy teaching behaviors A and B)
(complex teaching behaviors C)
A
B
C
Y
Zl
(less competent teaching Y)
(more competent teaching)

Crucial to the understanding of Guttman errors is that if teacher Y dominates the more complex teaching practice B, she or he should also dominate the less complex practice A. When teacher Y dominates the more complex teaching practice B, but not the less complex practice A, then it counts as one Guttman error. So, Guttman errors are not specific to an item, but specific to an item pair. They describe the situation where one teacher dominates a more complex teaching practice, but is dominated by an easy teaching practice, such that for this particular teacher the item order should be reversed. GNORMED counts the number of Guttman errors for each individual teacher and then divides the number of Guttman errors by the total possible number. This provides a coefficient which indicates the percentage of Guttman errors from .00 – 1.00. GNORMED of 1.00 indicates that the teacher shows all complex teaching practices while no easy teaching practices and GNORMED is .50 indicates that the teacher randomly shows some easy and some complex teaching practices (see for examples Table 6.1). To identify teachers who substantially deviate the cut-off criterion of GNORMED > .30 was used. That is, teachers showing more than 30% of the total possible number of Guttman errors are identified as teachers whom show a substantially different cumulative ordering.

Table 6.1

An example of five teacher-specific scoring patterns

	item 1	item 2	item 3	item 4	item 5	item 6	**Valid**	**GNORMED**
Teacher A	1						**Yes**	.00
Teacher B	1	1	1				**Yes**	.00
Teacher C	1	1	1	1	1		**Yes**	.00
Teacher …								
Teacher Y	1		1			1	**No**	.44
Teacher Z				1	1	1	**No**	-1.00

6.3.4 Software

The R package PerFit (Tenderio, 2014) is used to estimate GNORMED. Note that a considerable disadvantage of GNORMED is that it cannot compute the number of Guttman errors in case of a missing value. Cases with missing values are list wise deleted. Therefore, it is decided to add the χ^2-test of person fit to validate the findings with GNORMED. The advantage of the χ^2-test is that it can be applied to estimate fit irrespective of the number of missing values. Disadvantages are that it is relatively insensitive to small violations and in addition its power is dependent on the number of items included in the instrument.

6.4 Results

Table 6.1 presents the overall cumulative ordering in teaching practices from least complex "shows respect for student in behavior and language" to most complex "adapts processing of subject matter to student differences".

Table 6.2

Ordering in effective teaching practices from least (top) to most (bottom) complex

stage	**teaching practice**	**b**	***SE*$_{(b)}$**
climate	shows respect for students in behavior and language	−2.35	.373
instruction	explains the subject matter clearly	−1.96	.326
climate	creates a relaxed atmosphere	−1.68	.295
climate	ensures mutual respect	−1.43	.273
climate	supports student self-confidence	−1.40	.274
instruction	gives well-structured lessons	−1.40	.273
management	ensures that the lesson runs smoothly	−1.34	.267
management	ensures effective class management	−1.21	.256
management	uses learning time efficiently	−1.01	.243
instruction	gives feedback to students	−.69	.225
instruction	encourages students to do their best	−.57	.217
management	checks during processing whether students are carrying out tasks properly	−.27	.209
instruction	involves all students in the lesson	−.27	.203
instruction	clearly explains teaching tools and tasks	−.31	.211
activation	asks questions that encourage students to think	−.31	.204
activation	encourages students to reflect on solutions	−.18	.200
activation	uses teaching methods that activate students	−.11	.196
activation	has students think out loud	−.02	.194
activation	provides interactive instruction	.25	.187
instruction	checks during instruction whether students have understood the subject matter	.33	.183
learning strategies	encourages students to apply what they have learned	.47	.182

--- continues next page ---

stage	**teaching practice**	***b***	***SE(b)***
activation	boosts the self-confidence of weak students	.49	.181
learning strategies	teaches students how to simplify complex problems	.98	.178
learning strategies	encourages students to think critically	1.12	.174
learning strategies	encourages the use of checking activities	1.45	.176
learning strategies	teaches students to check solutions	1.35	.177
learning strategies	asks students to reflect on approach strategies	1.70	.178
differentiation	checks whether the lesson objectives have been achieved	1.72	.175
differentiation	adapts instruction to relevant student differences	2.21	.179
differentiation	offers weak students additional learning and instruction time	2.22	.180
differentiation	adapts processing of subject matter to student differences	2.23	.181

In general, most teachers dominate the descriptions at the top of the Table (i.e. most teachers show these practices), while the items at the bottom of the Table dominate most of the teachers (i.e. for most teachers these practices are too complex to show). In addition, the b-coefficient also provides estimates about the extent to which one teaching practice is more complex compared to another practice. Some practices cluster around similar b-values – suggesting that they are almost equally complex – while at some points the measurement scale shows 'gaps': for example, the item "uses learning time efficiently" ($b = -1.01$) is considerably less complex than "gives feedback to students" ($b = -.69$). Note further that the ordering in Table 6.2 is cumulative and that teachers showing practices located at the middle of the ordering will most likely also have performed most of the practices above 'the middle'. Also, the Table shows that – with some notable overlap – the cumulative ordering in practices follows the predicted six stages

The ordering presented in Table 6.2 presumes that the items and persons can be presented on a one-dimensional scale. Previous studies have provided evidence that observations of teaching practices fit this assumption (Van de Grift, Helms-Lorenz & Maulana, 2014). Using Guttman's (1954) simplex factor analysis (SFA) – which is specifically developed to estimate fit of cumulative data patterns – it is briefly evaluated again whether the assumption of one-dimensionality holds for this particular sample. To estimate the SFA, the Circum software is used (Browne, 1992). Note that the Circum specifies an additional constraint that items have similar distances on the measurement scale – which for our purposes is overly strict, but cannot be removed. The root mean square of error approximation (RMSEA) is .080. Given that an index between .05 and .08 reflects modest model fit (Hu & Bentler, 1999), the result indicates modest fit with the one-dimensional ordering. Given that 'modest' fit is sufficient for the current study purposes and given that this study is not meant to further validate the instrument, there was no further inspection on item (mis)fit.

6.4.1 An exploration of 'exceptional situations'

This study aims to investigate whether and how many teachers show practices which deviates from the hypothesized six stages in teacher development. For example teachers who perform most practices located in stage 3, while not performing many of the practices located in stage 1 or 2 show substantial deviations from our predictions. Thus, exceptional lesson situations are defined by a substantially different ordering and are explored using person fit statistics; in particular G_{NORMED} and the χ^2-statistic. A G_{NORMED} value > .30 indicates person misfit and likewise a significant χ^2-statistic indicates person misfit. In Table 6.3 the number of teachers who showed substantial deviations from the cumulative ordering are displayed (see Table 6.2 for the cumulative ordering).

Table 6.3

Number of teachers showing a substantial different ordering in teaching practices than the ordering provided in Table 6.2

	Valid cases	**Number showing different ordering**	**Percentage**
G_{NORMED} (>.30)	141	21	14.7
χ^2	198	31	15.7

Both G_{NORMED} > .30 and χ^2-test the identified the same cases as deviating from the baseline cumulative ordering, but the χ^2-test appeared slightly more lenient. G_{NORMED} > .30 identified four cases more than the χ^2-test (note that the χ^2-test identifies more cases because it also includes cases with missing values). Both statistics suggest that 15% of the lesson situations is exceptional and shows substantial deviations from the predicted cumulative ordering. It follows that these exceptional situations need to be identified and removed to avoid biased evaluation results.

6.4.2 Are exceptional lessons situations more typical to specific teachers?

An important question now is whether these 15% of lesson situations are incidents that may bias the evaluation of any teacher, or whether they are clustered around some specific teachers. The sample contained three observations of lesson situations for every teacher. This makes it possible to explore whether deviations occur more often with some specific teachers. Table 6.4 gives an overview how the exceptional lessons are distributed across teachers, observers, subjects, and classes. The Table gives the impression that exceptional lessons are not typically taught by specific teachers. In the column frequency (*f*) it can be observed that only three teachers repeatedly show a deviating ordering in teaching practices. In the column % it can be seen that these two lessons entail 66% of all lesson visits. These three teachers also had one third lesson observation which did fit the cumulative ordering.

When exceptional lessons would cluster around observers, this would suggest observer bias. Some observers tend to interpret teaching practices in ways which do not fit the overall interpretation. Again, however, the Table 6.4 suggest that observer bias is rather small. Observers who repeatedly evaluated teachers different from the cumulative ordering also had many other observations which did fit the ordering. Furthermore, exceptional lessons seem not clustered around specific subjects. Most lesson observations are concentrated within four different subjects: English as a foreign language (EFL) (*n* = 40), Dutch (as a first language) (*n* = 40), history (*n* = 42) and math (*n* = 44). The lessons Dutch deviate least often from the regular cumulative ordering, while the lessons history deviate most often. However, the overall differences between subjects seem to be small to negligible.

Maybe the strongest indicator of exceptional lesson situations is the class taught. Again, in no class more than half of all observations show deviations from the cumulative ordering. Yet, out of the 21 lesson situations identified as distinctively ordered, 15 concentrated among five classes out of the 23 classes in total. These five classes are taught by different teachers and observed by different observers, but multiple lessons show deviations from the cumulative ordering.

Table 6.4

The distribution of ‘exceptional’ lessons per background variable: the frequency (f) and percentage (%) relative to the total number of lesson observations

teacher			**observer**			**subject**			**class**		
ID	*f*	%	**ID**	*f*	%	**type**	*f*	%	**ID**	*f*	%
71	2	.66	66	2	.17	History	7	.17	202	4	.42
76	1	.50	68	4	.36	Math	5	.11	101	3	.33
158	1	.33	71	1	.25	English (EFL)	5	.13	37	4	.33
163	1	.50	78	1	.33	Dutch	2	.05	66	3	.43
170	1	.33	91	1	1.0	economics	1	.33	36	2	.17
172	1	.33	165	1	.33	Tech. drawing	1	.33	62	1	.08
173	1	.33	166	1	.33				32	1	.08
175	2	.66	208	1	.14				31	1	.08
176	1	.33	218	2	.20				61	1	.33
206	1	.33	219	1	.17				33	1	.11
210	1	.33	220	1	.25						
211	1	.33	221	1	.25						
212	1	.33	240	2	.66						
213	1	.33	242	2	.66						
215	1	.33									
240	1	.33									
241	2	.66									
242	1	.33									

Of these five classes, three concerned students in preparatory vocational education and two concerned students in higher preparatory vocational education. The sample comprised six classes from preparatory vocational education in total and 11 classes from higher preparatory vocational education. Note further that of the six classes concerning students in preparatory university education, only two classes each counted one lesson which deviated from the cumulative ordering. Based on this it may be speculated that classes in preparatory vocational education show more often deviations. However, also regarding these five special classes most lesson situations fit the ordering.

6.5 Conclusion

This study explored the possibility of individual differences in teachers' development of effective teaching practices. The study is grounded in a working theory predicting that lesson observation of teaching practices can be ordered using six stages. Previous studies have not been able to falsify this working theory, which implies that *most* teachers may indeed follow these six stages of development. However, if an individual teacher shows substantial deviations from the predicted six stages, it is evident that this teacher should be evaluated using an alternative working theory. Using person fit statistics, this study indicates that approximately 15% of the lesson situations show such substantial deviations. Further exploration indicates that misfit of the stage ordering seems not to be specific to some teachers. Only three teachers repeatedly deviated from the stage ordering on two (out of three) lesson situations observed. This provides few evidence to claim for individual differences in the development of teaching.

6.5.1 Limitations

Interpretation of the study results is restricted by the lack of a working theory for those lesson situations that do not fit the six stages. Most theoretical models lack an understanding of why persons or variables do not fit. However, if studies explore the characteristics of these particular deviating cases the lack of a guiding logic becomes more evident. Furthermore, the sample size of the study is limited, counting merely 69 teachers. If individual differences in teachers' development of teaching are truly exceptional, then the dataset may just have contained too few teachers to find them. This evidently restricts the implications of the results for theory on teacher development.

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.

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.

References

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38*, 123–140.

Antoniou, P., Kyriakides, L., & Creemers, B. P. M. (2015). The Dynamic Integrated approach to teacher professional development: rationale and main characteristics. *Teacher development, 19*(4), 535-552. doi: 10.1080/13664530.2015.1079550

Bachman, L. F. (2002). Alternative interpretations of alternative assessments: Some validity issues in educational performance assessments. *Educational Measurement: Issues and Practice*, *21*(3), 5-18.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear Mixed-Effects Models Using S4 Classes*. R package version 1.1-7. URL: http://CRAN.R-project.org/package=lme4.

Berliner, D. (2001). Learning about learning from expert teachers. *International Journal of Educational Research, 35,* 463–483.

Benton, S. L., & Cashin, W. E. (2012). *Student ratings of teaching: a summary of the research and literature.* (IDEA Paper No. 50). Retrieved March 3, 2015, from http://www.ntid.rit.edu/sites/default/files/academic_affairs/Sumry%20of%20Res%20%2350%20Benton%202012.pdf.

Borko, H. (2004). Professional Development and Teacher Learning: Mapping the Terrain, *Educational Researcher, 33*, 3-15. doi: 10.3102/0013189X033008003

Bill & Melinda Gates Foundation (2012). *Asking students about teaching: Student perception surveys and their implementation.* Retrieved March 3 from http://www.metproject.org/downloads/Asking_Students_Practitioner_Brief.pdf.

Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences.* London: Lawrence Erlbaum Associates.

Brennan, R. L. (2001). *Generalizability theory: Statistics for social science and public policy.* New York: Springer-Verlag.

Brennan, R. L. (2004). *Some perspectives on inconsistencies among measurement models.* (CASMA Research Report No. 10). Retrieved March 9, 2015, from http://www.uiowa.edu/~casma/NSF-casma-rpt.pdf.

Briggs, D. C., & Wislon, M. (2007). Generalizability in Item response theory. *Journal of Educational Measurement, 44*, 131 – 155.

Browne, M. W. (1992). Circumplex models for correlation matrices. *Psychometrika, 57,* 469-497.

Cai, L., Thissen, D., & Du Toit, S. (2005–2013). IRTPRO (Version 2.1) [Computer software]. Lincolnwood, IL: Scientific Software International.

Centra, J. A. (1975). Colleagues as raters of classroom instruction. *The Journal of Higher Education, 46*(3)*,* 327-337. doi: 10.2307/1980806

Chen, W-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265-289.

Cheung, W. M., & Wong, W. M. (2013). Does lesson study work? A systematic review on the effects of lesson study and learning study on teachers and students. *International Journal for Lesson and Learning Studies, 3*(2), 137-149. DOI 10.1108/IJLLS-05-2013-0024

Choi, J. (2013). *Advances in combining generalizability theory and item response theory.* Doctoral dissertation, University of California, Berkeley.

Conway, P. F., & Clark, C. M. (2003). The journey inward and outward: A re-examination of Fuller's concerns-based model of teacher development. *Teaching and Teacher Education, 19*, 465–482.

CPS, (2012). *Ziet u het verschil? Ook in het voortgezet onderwijs is differentiëren essentieel* [*Do you see the difference? The relevance of differentiation in secondary education*] [special issue]. Didactief, 42(8).

Creemers, B. P. M., & Kyriakides, L. (2005) Critical analysis of the current approaches to modelling educational effectiveness: The importance of establishing a dynamic model. *School Effectiveness and School Improvement, 17*(3), 347-366. doi: 10.1080/09243450600697242

Cromley, J. G., Perez, T. C., Fitzhugh, S. L., Newcombe, N. S., Wills, T. W., & Tanaka, J. C. (2013). Improving students' diagram comprehension with classroom instruction. *Journal of experimental education, 81,* 511-537. doi: 10.1080/00220973.2012.745465

Cronbach, L., J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-333.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer, & H. I. Braun (Eds.), *Test validity* (pp 3-17). New Jersey, USA: Lawrence Erlbaum Associates Inc.

Cronbach, L. J. (2004). My Current Thoughts on Coefficient Alpha and Successor Procedures. *Educational and psychological measurement, 64,* 391-418. doi: 10.1177/0013164404266386

Cronbach, L., J. & Meehl, P., E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281-302

Cronbach, L. J., Gleser, C. G., Rajaratnam, N., & Nanda, H. (1972). *The dependability of behavioral measurements.* New York, USA: Wiley.

Dall'Alba, G., & Sandberg, J. (2006). Unveiling Professional Development: A Critical Review of Stage Models. *Review of educational research, 76*(3), 383-412.

Danielson, C. (2013). *The framework for teaching: Evaluation instrument.* Princeton, NJ: The Danielson Group.

Darling-Hammond, L. (2013). *Getting teacher evaluation right. What really matters for effectiveness and improvement.* New York, USA: Teachers College Press

Darling-Hammond, L., Amrein-Beardsley, A., Heartel, E., & Rothstein J. (2012). Evaluation teacher evaluation. *Phi Delta Kappan, 93,* 8–15.

Day, C., Sammons, P., Stobart, G., Kingston, A., & Gu, Q. (2007). *Teachers matter: Connecting lives, work and effectiveness.* Maidenhead, UK: Open University Press.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Abe, H., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software, 39,* 1–25.

De Jong, R., & Westerhof, K.J. (2001). The quality of student ratings of teacher behaviour. *Learning Environments Research 4,* 51–85.

DeMars, C. (2010). *Item response theory: Understanding statistics measurement.* New York: Oxford University Press.

DfEE (2012). *Teacher appraisal and capability. A model policy for schools.* Retrieved June 30, 2015, from: https://www.gov.uk/government/uploads/system/uploads/ attachment_data/file/282598/Teacher_appraisal_and_capability.pdf

Doran, H., Bates, D., Blies, P., & Dowling, M. (2007). Estimating the multilevel Rasch model with the lme4 Package. *Journal of Statistical Software, 20*, 1-18.

Doyle, W. (1983). Academic work. *Review of Educational Research, 53*(2), 159-199.

Doyle, W. (2006). Ecological approaches to classroom management. In: C.M. Evertson & C.S. Weinstein (Eds), *Handbook of classroom management*: *research, practice, and contemporary issues* (p. 97- 125). New York: Erlbaum.

Doyle, W. (2009). Situated Practice: A Reflection on Person-Centered Classroom Management. *Theory Into Practice, 48*(2), 156-159, doi: 10.1080/00405840902776525

Ebbens, S., & Ettekoven, S. (2005). *Effectief leren: Basisboek* [*Effective learning: Basic handbook*]. Groningen, The Netherlands: Wolters-Noordhoff.

Feldman, K. A. (1989). Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external (neutral) observers. *Research in Higher Education, 30*(2), 137-194.

Firestone, W. A. (2014). Teacher evaluation policy and conflicting theories of motivation. *Educational Researcher 43,* 100-107.

Fox, J-P. (2010). *Bayesian Item Response Modeling. Theory and Applications.* New York: USA, Springer

French-Lazovik, G. (1981). Documentary evidence in the evaluation of teaching. In J. Millman (Ed.), *Handbook of Teacher Evaluation*. Beverly Hills, CA: Sage Publications.

Fuller, F. (1969). Concerns of teachers: A developmental conceptualization. *American Educational Research Journal, 6*, 207–226.

Gelman A., Su Y-S, Yajima M, Hill J, Pittau M. G., Kerman J., et al. *Arm: Data analysis using regression and multilevel/ hierarchical models*. R package version 1.8-6 2015: URL: https://cran.r-project.org/web/packages/arm/arm.pdf.

Goldhaber, D., & Hansen, M. (2013). Is it just a bad class? Assessing the long-term stability of teacher performance. *Economica, 80*, 589–612. doi:10.1111/ecca.12002

Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The test matters: The relationship between classroom observation scores and teacher value added on multiple types of assessment. *Educational Researcher, 43,* 293 – 303. DOI:10.3102/0013189X14544542

Guilleux, A., Blanchin, M., Hardouin, J-B., Sébille, V. (2014). Power and Sample Size Determination in the Rasch Model: Evaluation of the Robustness of a Numerical Method to Non-Normality of the Latent Trait. *Plus One, 9*(1), 1-7.

Guttman, L. L. (1954). A new approach to factor analysis: the radex. In Lazersfeld, Paul F. (Ed.), *Mathematical thinking in the social sciences*. Glencoe, Illinois: The Free Press

Guttman, L. L. (1977). What is not what in statistics. *Journal of the Royal Statistical Society, 26*(2), 81-107.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Abingdon, Oxon, UK: Routledge.

Haberman, S. J. (2008) When can subscores have value? *Journal of Educational and Behavioral Statistics, 33*(2), 204-229. doi: 10.3102/1076998607302636

Hanushek, E. A. (2007) . The single salary schedule and other issues of teacher pay. *Peabody Journal of Education, 82*, 574-586. doi: 10.1080/01619560701602975

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review, 30*(3), 466-479. doi: 10.1016/j.econedurev.2010.12.006

Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A. J., Jones, S. M., Brown, J. L., Cappella, E., Atkins, M., Rivers, S. E., Brackett, M. A., & Hamagami, A. (2013). Teaching through interactions: testing a developmental framework of teaching effectiveness in over 4,000 classrooms. *The Elementary School Journal, 113,* 461-487.

Hazi, H. M., & Rucinsky, D. A. (2009). Teacher Evaluation as a Policy Target for Improved Student Learning: A Fifty-State Review of Statute and Regulatory Action since NCLB. *Educational Policy Analysis Archives, 17*(5), 1-22.

Helms-Lorenz, M., Van de Grift, W. J. C. M., & Maulana, R. (2016). Longitudinal effects of induction on teaching skills and attrition rates of beginning teachers. *School Effectiveness and School Improvement, 27*(2), 178-204. Doi: 10.1080/09243453.2015.1035731

Hill, H. C., Besiegel, M., & Jacob, R. (2013). Professional development research: Consensus, crossroads and challenges. *Educational Researcher, 42,* 476-487. doi: 10.3102/0013189X13512674

Hill, H., Charalambous, C. Y., & Kraft, M. A. (2012). When interrater-reliability is not enough: Teacher observation systems and a case for the generalizability theory. *Educational Researcher 41,* 56–64. doi: 10.3102/0013189X12437203.

Hill, H., Kapitula, L., & Umland, K. (2011). A validity argument to evaluating teacher value added scores. *American Educational Research Journal 48*, 797–831.

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel.* Seattle, WA: Bill & Melinda Gates Foundation.

Howard, G. S., Conway, C. G., & Maxwell, S. E. (1985). Construct validity of measures of college teaching effectiveness. *Journal of Educational Psychology, 77*(2), 187-196.

Hoxby, C. M. (2002). *The cost of accountability* (No. w8855). National Bureau of Economic Research.

Hox, J. (2010). Multilevel analysis: Techniques and applications (2nd edition). New York, NY: Routledge.

Hu, L., & Bentler, M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55. doi: 10.1080/10705519909540118

Huberman, M. (1993). *The lives of teachers.* New York, NY: Teachers College Press.

Inspectie van het Onderwijs (2009). *International Comparative Analysis of Learning and Teaching in Math Lessons in Several European Countries.* De Meern, Inspectie van het Onderwijs.

Inspectie van het Onderwijs (2016). *De staat van het onderwijs 2014-2015* [*The state of education in the Netherlands 2014-2015*]. De Meern, Inspectie van het Onderwijs.

Inspectie van het Onderwijs (2015). *De staat van het onderwijs 2013-2014* [*The state of education in the Netherlands 2013-2014*]. De Meern, Inspectie van het Onderwijs.

Isoré, M. (2009). *Teacher Evaluation: Current Practices in OECD Countries and a Literature Review.* OECD Education Working Papers, No. 23. OECD Publishing (NJ1).

Jöreskog, K. G. (1970). Estimation and testing of simplex models. *The British Journal of Mathematical and Statistical Psychology, 23*(2), 121-145.

Kagan, D. M. (1992). Professional growth among pre-service and beginning teachers. *Review of Educational Research, 62*, 129–169.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington, DC: The National Council on Measurement in Education & the American Council on Education.

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement, 50*(1), 1-73. doi: 10.1111/jedm.12000

Kane, T. J., Staiger, D. O., McCaffrey, D., Cantrell, S., Archer, J., Buhayar, S., Kerr, K., Kawakita, T., & Parker, D. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Seattle, WA: Bill & Melinda Gates Foundation.

Kennedy, M. (2010). Attribution error and the quest for teacher quality. *Educational researcher, 39*, 591-598. doi: 10.3102/0013189X10390804

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (Third edition). New York: Guilford Press.

Koller, I. & Hatzinger R. (2013). Nonparametric tests for the Rasch model: explanation, development, and application of quasi-exact tests for small samples. *Interstat, 11,* 1-16.

Kyriakides, L. (2013). What matters for student learning outcomes: A meta-analysis of studies exploring factors of effective teaching. *Teaching and Teacher Education, 36*, 143–152.

Kyriakides, L., Creemers, B. P. M., & Antaniou, P. (2009). Teacher behavior and student outcomes: Suggestions for research on teacher training and professional development. *Teaching and Teacher Education, 25,* 12–23.

Louws, M. (2016). *Professional learning: What teachers want to learn* (Doctoral Dissertation). Leiden, the Netherlands: ICLON.

Mainhard, T. M., Brekelmans, M., Den Brok, P., & Wubbels, T. (2011). The development of the classroom social climate during the first months of the school year. *Contemporary Educational Psychology, 36*(3), 190-200. doi:10.1016/j.cedpsych.2010.06.002

Mair, P., & Hatzinger, R. (2007). Extended Rasch modelling: The eRm package for the application of IRT models in R. *Journal of Statistical Software, 20*, 1–20.

Marsh, H. D. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–383). Dordrecht, The Netherlands: Springer.

Marzano, R.J. (2003). *What works in schools: Translating research into action.* Alexandria, VA: Association for Supervision and Curriculum Development.

Marzano, R. J. (2012). The two purposes of teacher evaluation. *Educational Leadership, 70,* 14-19.

Marzano, R. J., & Toth, M. D. (2013). *Teacher evaluation that makes a difference. A new model for teacher growth and student achievement.* Alexandria, Virginia: ASCD

Maualan, R., & Helms-Lorenz, M. (2016). Observations and student perceptions of the quality of preservice teachers' teaching behaviour: construct representation and predictive quality. *Learning environments research, 19*(3), 335-357. doi:10.1007/s10984-016-9215-8

Maulana, R., Helms-Lorenz, M., & Van de Grift, W. (2015). Development and evaluation of a questionnaire measuring pre-service teachers' teaching behaviour: A Rasch modelling approach. *School Effectiveness and School Improvement, 26*(2), 169-194.

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement, 18*(4)*,* 311–314.

Meijer, R. R., Sijtsma, K., & Smid, N. G. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. *Applied Psychological Measurement, 14*(3), 283-298.

Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013). *A composite estimator of effective teaching.* Seattle, WA: Bill & Melinda Gates Foundation.

Mourshed, M., Chijioke C., & Barber, M. (2010). *How the world's most improved school systems keep getting better.* London: McKinsey Company.

Muijs, D. (2006). Measuring teacher effectiveness: Some methodological reflections. *Educational Research and Evaluation: An International Journal on Theory and Practice 12,* 53–74.

Muijs, D., Kyriakides, L., van der Werf, G., Creemers, B., Timperley, H., & Earl, L. (2014). State of the art—teacher effectiveness and professional learning, *School Effectiveness and School Improvement, 25*(2), 231–256, doi: 10.1080/09243453.2014.885451

Murray, H. G. (1983). Low-inference classroom teaching and student ratings of college teaching effectiveness. *Journal of Educational Psychology, 75*(1), 138-149.

Murillo, F. J. (2007). *Evaluación del desempeño docente y carrera profesional docente. Un estudio comparado entre 50 países de América y Europa.* Santiago de Chile: OREALC/UNESCO.

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (7th ed.). Los Angeles: Muthén and Muthén.

NCTQ (2013). *Connect the dots: Using evaluations of teaching effectiveness to inform policy and practice.* Washington, DC: National Council on Teacher Quality.

Nusche, D., Braun, H., Halász, G., & Santiago, P. (2014). *OECD Reviews of Evaluation and Assessment in Education: Netherlands 2014.* OECD Reviews of Evaluation and Assessment in Education, OECD Publishing.

http://dx.doi.org/10.1787/9789264211940-en

Nunnally, J. C. (1978). *Psychometric theory.* New York: McGraw-Hill.

OCW [Dutch Ministry of Education, Culture, and Science] (2013a). *Peer review in de praktijk* [*Peer review in practice*]. Rotterdam: VOION.

OCW [Dutch Ministry of Education, Culture, and Science] (2013b). *Begeleiding van beginnende leraren in het beroep* [*Induction of inexperienced teachers*]. The Hague, The Netherlands: OCW: http://www.leroweb.nl/cms/wp-content/uploads/2013/11/Begeleiding-beginnende-leraren.pdf

OECD (2016), *Netherlands 2016: Foundations for the Future*, OECD Publishing, Paris. doi: http://dx.doi.org/10.1787/9789264257658-en

Overdiep I. (2016). *Wijzer over Zien en Kijken: Inventarisatie observatie instrumenten in het PO* [*learning from observation: a review of observation instruments applied in Primary Education*]. Utrecht, The Netherlands: PO-raad. https://www.poraad.nl/files/werkgeverszaken/wijzer_over_zien_en_kijken.pdf

Patrick, H., & Mantzicopolous, P. (2016). Is effective teaching stable? *Journal of Experimental Education, 84,* 23-47. doi: 10.1080/00220973.2014.952398

Patz, R. P., Jucker B. W., Johnson, M. S., & Mariano, L.T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics, 27,* 341–384. doi: 10.3102/10769986027004341

Peterson, K. D. (2000). *Teacher evaluation: A comprehensive guide to new directions and practice.* Thousand Oaks, CA: Corwin Press.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: standardized observation can leverage capacity. *Educational researcher, 38*(2), 109-119. doi: 10.3102/0013189X09332374

Ponocny, I. (2001). Non-parametric goodness-of-fit tests for the Rasch model. *Psychometrika 66*, 437–460.

Popham, (1988). The dysfunctional marriage of formative and summative teacher evaluation. *Journal of Personnel Evaluation in Education, 1,* 269 – 273.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Nielsen & Lydiche.

Raju, N. S., Price, L. R., Oshima, T. C., & Nering, M. L. (2006). Standardized conditional SEM: A case for conditional reliability. *Applied Psychological Measurement, 30,* 1–12. doi: 10.1177/0146621606291569.

Richardson, J. T. E. (2005). Instruments for obtaining student feedback: A review of the literature. *Assessment and Evaluation in Higher Education, 30*, 387–415. doi: 10.1080/02602930500099193.

Richardson, V., & Placier, A. (2001). Teacher change. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed.). Washington, DC: American Educational Research Association.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software, 17*, 1–25.

Schafer, E., Stringfield, S., & Wolfe, D. (1992). Two-year effects of a sustained beginning teacher induction program on classroom interactions. *Journal of Teacher Education, 43,* 203–214.

Shepard, L. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of research in education* (pp. 405–450). Washington, DC: American Educational Research association.

Scriven, M. (1981). Summative teacher evaluation. In J. Millman (Ed.), *Handbook of Teacher Evaluation*. Beverly Hills, CA: Sage Publications.

Shavelson, R., & Dempsey-Atwood, N. (1976). Generalizability of teaching behavior. *Review of Educational Research, 46,* 553-611.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability Theory: A Primer.* Thousand Oaks, California: Sage Publications, Inc

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, *57*(1), 1-23.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*(1), 72-101.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 271-295.

Steffy, B. E., & Wolfe, M. P. (2001). A life-cycle model for career teachers. *Kappa Delta Pi Record 38*, 16–19. doi: 10.1080/00228958.2001.10518508.

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Difference, 42*, 893-898.

Sternberg, R., J. & Horvath, J. A. (1995). A prototype view of expert teaching. *Educational Researcher, 24*, 9–17.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. Psychometrika, 55(2), 293-325.

Strong, M. (2011). *The high qualified teacher. What is teacher quality and how do we measure it?* NY, NY: Teachers College press

Teitler, P. (2013). *Lessen in orde* [*Lessons in order*]. Bussum, The Netherlands: Uitgeverij Couthino

Tendeiro, J. N. (2014). Package 'PerFit' (published online). In R. Cran (Ed.), *The comprehensive R network.* retrieved from: http://cran.r-project.org/web/packages/PerFit/PerFit.pdf.

Thorndike, E. L. (1918). Individual differences. *Psychological Bulletin*, *15*, 148–159.

Timmerman, M. E., Lorenzo-Seva, U. & Ceulemans, E. (in press). The number of factors problem. in P. Irwing, T. Booth, & D.J. Hughes. (eds.; in press). *The Wiley Handbook of Psychometric Testing*, John Wiley & Sons, Chichester, UK.

Toch, T., & Rothman, R. (2008). *Rush to judgment: Teacher evaluation in public schools.* Washington DC: Education Sector.

U.S. Department of Education (2009). *Race to the Top Program: Executive Summary*. Washington, USA.

Vale, C. D. (1986). Linking Item parameters onto a common scale. *Applied Psychological Measurement, 10*(4), 333-344.

Van Veen, K., Zwart, R., Meirink, J., & Verloop, N. (2010). *Professionele ontwikkeling van leraren* [*Professional development of teachers*]. Leiden, The Netherlands: ICLON: Expertisecentrum leren van docenten.

Van de Grift, W. (1990). Het onderzoek naar effectieve scholen. *Pedagogische Studiën*, *67* (10) 462-463

Van de Grift, W. J. C. M. (2007). Quality of teaching in four European countries: a review of the literature and application of an assessment instrument. *Educational Research, 49*(2), 127-152. doi: 10.1080/00131880701369651

Van de Grift, W. J. C. M. (2013). *Van zwak naar sterk. De aanpak van zwakke en zeer zwakke scholen voor voortgezet onderwijs in het noorden van het land door* *observatie van en feedback voor leraren*. Drachten: The Netherlands. https://www.rug.nl/staff/w.j.c.m.van.de.grift/vanzwaknaarsterkdrachten.pdf

Van de Grift, W. J. C. M. (2014). Measuring teaching quality in several European countries. *School effectiveness and school improvement, 25*(3), 295–311 doi: 10.1080/09243453.2013.794845

Van de Grift, W. J. C. M. & J.F. Lam (1998). Het didactisch handelen in het basisonderwijs. [Teachers' instructions in primary education.] *Tijdschrift voor Onderwijsresearch 23*(3), 224-241.

Van de Grift, W. J. C. M., Helms-Lorenz, M., & Maulana, R. (2014). Teaching skills of student teachers: Calibration of an evaluation instrument and its value in predicting student academic engagement. *Studies in Educational Evaluation, 43*, 150–159 doi: 10.1016/j.stueduc.2014.09.003

Van de Grift, W. J. C. M., Van der Wal, M., & Torenbeek, M. (2011). Ontwikkeling in de pedagogische didactische vaardigheid van leraren in het basisonderwijs. [Primary teachers' development of pedagogical didactical skill.] *Pedagogische Studiën, 88*, 416–432.

Van der Ark, A., L. (2007). Mokken Scale Analysis in R. *Journal of Statistical Software, 20*(11), 1-19. doi: 10.18637/jss.v020.i11

Van der Lans, R. M., Van de Grift, W. J., & Van Veen, K. (2015). Developing a Teacher Evaluation Instrument to Provide Formative Feedback Using Student Ratings of Teaching Acts. *Educational Measurement: Issues and Practice, 34*(3), 18-27.

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

Van der Lans, R. M., Van de Grift, W. J. C. M., Van Veen, K., & Fokkens-Bruinsma, M. (2016). Once is not enough: Establishing reliability criteria for feedback and evaluative decisions using classroom observation. *Studies in Educational Evaluation, 50*, 88-95.

Van Veen, K. (2011). Het niveau en de kwalitateit van leraren in het basisonderwijs en voortgezet onderwijs: wat is het probleem? [the competence and quality of primary and secondary education teachers: what is the problem?] *Pedagogische studiën*, 433-441.

Van Veen, K., Zwart, R., Meirink, J., & Verloop, N. (2010). *Professionele ontwikkeling van leraren: Een reviewstudie naar effectieve kenmerken van professionaliseringsinterventies van leraren* [*Teacher professional development: A review study on effective characteristics of teacher professional development interventions*]. Leiden: ICLON / Expertisecentrum Leren van Docenten [ICLON / Expertise centre Teacher learning].

Van Schuur, W. H. (2011). *Ordinal Item response theory: Mokken scale analysis* (Vol 169). Thousand Oaks, California: Sage publications

Vitikka, E., Krokfors, L., & Hurmerinta, E. (2012). *The Finnish National Core Curriculum: Structure and development* (draft). Retrieved September 2016 from: http://curriculumredesign.org/wp-content/uploads/The-Finnish-National-Core-Curriculum_Vitikka-et-al.-2011.pdf

Von Davier, M., & Carstensen, C. H. (2007). *Multivariate and mixture distribution Rasch models: extensions and applications.* Springer: New York.

Weeks, J. P. (2010). plink: A R Package for linking mixed-format tests using IRT-Based methods. *Journal of Statistcal Software 35*(12), 1-33. doi: 10.18637/jss.v035.i12.

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., & Morgan, K. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness.* New Teacher Project.

Winters, M. A., & Cowen, J. M. (2014). Who would stay, who would be dismissed? An empirical consideration of value-added teacher retention policies. *Educational Researcher, 42,* 330–337. DOI: 10.3102/0013189X13496145

Wubbels, T., & Brekelmans, M. (2005). Two decades of research on teacher-student relationships in class. *International Journal of Educational Research, 43,* 6-24. doi:10.1016/j.ijer.2006.03.003

Samenvatting

Algemene introductie

De leraar speelt een grote rol in het leren van leerlingen. Mede om deze reden leggen politici en beleidsmakers steeds meer nadruk op het evalueren van leraren als middel om het onderwijs te verbeteren. Evaluatie dient daarin twee doelen. Ten eerste zou evaluatie moeten leiden tot feedback waarmee leraren hun lesgeven kunnen verbeteren. Ten tweede zou evaluatie moeten bijdragen aan besluiten over leraren met betrekking tot bijvoorbeeld salarisschalen, contractverlening of, in extreme gevallen, ontslag. Het geven van feedback en het nemen van zulke evaluatieve besluiten is niet nieuw. Alle scholen moeten immers personeelsbeleid voeren. Nieuw in het beleid rondom evaluatie van leraren is de relatief grote nadruk op hun bekwaamheid in lesgeven en ook nieuw is de hogere frequentie waarin dit zou moeten worden geëvalueerd (het streven is dat iedere docent jaarlijks een functioneringsgesprek heeft[3]). Met deze nieuwe focus is het belangrijk geworden om de bekwaamheid in lesgeven valide en betrouwbaar te evalueren. Wanneer evaluaties onbetrouwbaar zijn dan is er een onacceptabel grote kans dat leraren verkeerde feedback ontvangen of dat er over hen verkeerde evaluatieve besluiten worden genomen. Verkeerde feedback zal bovendien niet leiden tot een verbetering in lesgeven.

In dit proefschrift wordt de validiteit en betrouwbaarheid onderzocht van de feedback aan leraren en de evaluatieve besluiten over leraren op basis van lesobservaties en leerlingenvragenlijsten. In lijn met de duale doelstelling van de evaluatie van leraren worden er twee verschillende vragen van validiteit behandeld. Om informatie uit lesobservaties en leerlingenvragenlijsten te kunnen gebruiken om leraren feedback te geven is het belangrijk om wetenschappelijk bewijs te vinden hoe de vaardigheid in lesgeven zich ontwikkelt. Alleen wanneer evaluatie van leraren gekoppeld is aan een algemene theorie die voorspelt hoe de vaardigheid van lesgeven van een leraar zich zal ontwikkelen kan de informatie over het lesgeven van een leraar worden gerelateerd aan een concreet advies voor verdere ontwikkeling. Daarom is in 3 verschillende steekproeven onderzocht hoe vaardigheid van lesgeven zich in het algemeen ontwikkelt bij docenten. Dit onderzoek naar de algemene ontwikkeling wordt beschreven in Hoofdstuk 2, 3, en 4. Om de informatie uit lesobservaties en leerlingenvragenlijsten te gebruiken in functioneringsgesprekken is het belangrijk wetenschappelijk aan te tonen of en onder welke voorwaarden deze informatie hiervoor betrouwbaar genoeg is. Goede besluitvorming vereist dat er voldoende informatie

[3] Zie bijvoorbeeld: https://www.delerarenagenda.nl/de-lerarenagenda/scholen-als-lerende-organisaties

is verzameld vanuit verschillende gezichtspunten (zodat falsificatie mogelijk is) en dat inaccurate informatie wordt opgemerkt en niet wordt meegewogen. Om deze reden worden in Hoofdstuk 5 en 6 vraagstukken rondom de betrouwbaarheid beschreven.

Vanuit bovenstaande is de volgende hoofdvraag geformuleerd die vervolgens is opgedeeld in 5 deelvragen.

Hoe kunnen lesobservaties en leerlingenvragenlijsten op zo'n manier worden ingezet in scholen dat ze leraren op een valide en betrouwbare manier van feedback kunnen voorzien en dat scholen op een valide en betrouwbare manier evaluatieve besluiten over leraren kunnen nemen?

Deelvragen:

1. Kunnen observaties van het pedagogisch en didactisch handelen van docenten tijdens de les stapsgewijs geordend worden, zodat minder complexe handelingen eerst getoond moeten worden alvorens meer complexe handelingen kunnen worden getoond? En: Hoe ziet deze stapsgewijze ordening eruit? **(Hoofdstuk 2)**
2. Kunnen leerlingenantwoorden op items die refereren naar het pedagogisch en didactisch handelen tijdens de les stapsgewijs geordend worden? En: Hoe ziet deze stapsgewijze ordering eruit? **(Hoofdstuk 3)**
3. In hoeverre is er overeenstemming tussen lesobservatoren en leerlingen in hoe zij de stapsgewijze ontwikkeling indelen? En: vinden lesobservatoren en leerlingen gelijkwaardig gedrag ongeveer even complex? **(Hoofdstuk 4)**
4. Hoeveel lesobservaties door lesobservatoren zijn noodzakelijk voor een bescheiden betrouwbaarheid, voldoende om te gebruiken voor feedback aan leraren? En hoeveel lesobservaties door lesobservatoren zijn noodzakelijk voor een hoge betrouwbaarheid, voldoende om te gebruiken om de informatie te gebruiken in functioneringsgesprekken? **(Hoofdstuk 5)**
5. Zijn er leraren die zich anders ontwikkelen dan de stapsgewijze volgorde beschreven in hoofdstuk 2, 3 en 4? En: Hoe kunnen we deze unieke gevallen herkennen? **(Hoofdstuk 6)**

Hoofdstuk 2 en 3

De hoofdstukken 2 en 3 richten zich op hoe de vaardigheid in lesgeven zich ontwikkelt. De hoofdstukken zijn gebaseerd op dezelfde theorie en in deze hoofdstukken worden ongeveer dezelfde methoden gebruikt. Daarom worden ze hier gezamenlijk samengevat.

Introductie

Onderwijsbeleid richt zich steeds meer op lerarenevaluatie als een middel om leraren feedback te geven over hun functioneren. Daarom is een aantal instrumenten ontwikkeld waarmee scholen de vaardigheid van lesgeven kunnen evalueren. Er is echter kritiek op deze instrumenten omdat alleen het terugkoppelen van de huidige vaardigheid in lesgeven leraren weinig informatie biedt over hoe ze zich kunnen verbeteren en doorontwikkelen. Net zoals leerlingen in stappen de lesstof leren lijkt het aannemelijk dat leraren leren lesgeven in stappen. Feedback zou daarom beter gericht kunnen worden op waar in de ontwikkeling de leraar zich nu bevindt en in welke pedagogische en didactische handelingen de leraar zich zou moeten bekwamen om een stap verder te komen.

Om meer inzicht in de stappen te verkrijgen wordt in de hoofdstukken 2 en 3 bestudeerd of observaties van het pedagogisch en didactisch handelen van de leraar tijdens de les stapsgewijs (cumulatief) kunnen worden geordend en of deze ordening overeenstemt met de eerdere bevindingen beschreven in theorie over de ontwikkeling van leraren. Onder observaties worden zowel lesobservaties door collega’s of inspecteurs verstaan (Hoofdstuk 2) als observaties door leerlingen (Hoofdstuk 3). Op basis van eerder onderzoek wordt de hypothese geformuleerd dat de pedagogische en didactische vaardigheid in lesgeven zich ontwikkelt in grofweg de volgende zes stappen: (1) veilig en stimulerend leerklimaat, (2) efficiënte lesorganisatie en (3) duidelijke en gestructureerde instructie, (4) intensieve en activerende les, (5) leerstrategieën aanleren en (6) afstemmen van instructie. Hierbij is de hypothese ook dat een zeker mate van bekwaamheid in de handelingen beschreven in een lagere stap voorwaardelijk is om bekwaam te worden in de handelingen beschreven in een hogere stap (bijvoorbeeld kundigheid in stap 1 is voorwaardelijk voor kundigheid in stap 2). Als deze hypothese klopt dan zouden observaties van lesgeven grofweg het patroon moeten volgen, zoals weergegeven in Figuur S1. In deze Figuur is een vinkje genoteerd in het geval dat een docent bekwaam is in de handelingen passend bij deze stap. Het patroon geeft weer dat sommige docenten zich al in meer stappen bekwaam tonen dan collega-docenten. Het is ook duidelijk dat geen van de docenten zich bekwaam toont in een hogere stap zonder dat ze zich ook bekwaam tonen in een lagere stap.

Figuur S1

De veronderstelde stapsgewijze ontwikkeling van leraren. Een vinkje betekent dat een leraar de (meeste) handelingen behorend bij deze stap voldoende beheerst.

	veilig en stimulerend klimaat	efficiënt les-organisatie	duidelijke en gestructureerde instructie	intensieve en activerende les	leerstrategieën aanleren	afstemmen van instructie
Docent A	✓					
Docent B	✓	✓				
Docent C	✓	✓	✓			
Docent D	✓	✓	✓	✓		
Docent E	✓	✓	✓	✓	✓	
Docent F	✓	✓	✓	✓	✓	✓

Methode

Voor het onderzoek in Hoofdstuk 2 zijn 878 lesobservaties uitgevoerd op 119 scholen verspreid over Nederland. Van alle observaties is 46.6% uitgevoerd door inspecteurs van de onderwijsinspectie, de overige 53.4% is uitgevoerd door collega-leraren. Alle observaties zijn uitgevoerd met hetzelfde instrument: de ICALT. Het ICALT-lesobservatie formulier bestaat uit 32 items (zie appendix D). De lesobservaties zijn uitgevoerd in het schooljaar 2010-2011. In de analyse wordt gebruik gemaakt van het Rasch-model dat specifiek ontwikkeld is om cumulatieve item volgorde te toetsen.

Voor het onderzoek in Hoofdstuk 3 zijn 1590 leerlingenvragenlijsten geanalyseerd die zijn ingevuld voor 68 leraren werkzaam op één school. Deze vragenlijsten zijn afgenomen in het schooljaar 2011-2012. De zogenoemde “Mijn Leraar” vragenlijst, die is gebruikt voor dit onderzoek bestond in een eerste versie uit 59 items. Items beschrijven een specifieke handeling van de leraar en de observator of leerling wordt gevraagd of de handeling in voldoende mate wordt uitgevoerd of niet (voor de specifieke instrumenten zie appendix D en E). Om te onderzoeken of we de pedagogische en didactische handelingen kunnen ordenen volgens een stapsgewijs patroon is gebruik gemaakt van het Rasch-model.

Resultaten

Uit de resultaten komt naar voren dat 31 van de 32 items van de ICALT passen binnen het stapsgewijs patroon en dat 32 van de 59 vragen op de "Mijn Leraar" vragenlijst passen binnen een stapsgewijs patroon. De volgorde van de handelingen komt grotendeels overeen met de hypothese van de zes stappen in ontwikkeling: (1) *veilig en stimulerend leerklimaat*, (2) *efficiënte lesorganisatie* (3) *duidelijke en gestructureerde instructie*, (4) *intensieve en activerende les*, (5) *leerstrategieën aanleren* en (6) *afstemmen van instructie*. Toch blijkt wel dat de stappen niet helemaal los van elkaar kunnen worden gezien. Het lijkt bijvoorbeeld niet zo te zijn dat een leraar zich eerst compleet moet bekwamen in alle handelingen behorend bij stap 2 voordat de leraar zich kan bekwamen in de pedagogische en didactische handelingen uit stap 3.

Conclusie & Implicaties

Zowel de ICALT-lesobservaties als de "Mijn Leraar" leerlingenvragenlijst bevestigen de hypothese. Daarmee bieden deze studies ondersteunend bewijs voor de stelling dat de ontwikkeling in pedagogische en didactische vaardigheid van leraren zich cumulatief ontwikkelt.

Met dit onderzoek is meer inzicht verkregen in de ontwikkeling in pedagogische en didactische vaardigheid van leraren. Eerdere inzichten waren meestal beperkt tot minder ervaren leraren (<3 jaar), terwijl dit onderzoek leraren van alle ervaringsjaren meegenomen heeft. Met de gevonden stapsgewijze ontwikkeling in pedagogisch-didactische handelingen is het bovendien mogelijk om preciezer feedback te geven aan leraren, zodat duidelijk kan worden aangegeven in welke pedagogische en didactische handelingen de leraar al vaardig is, welke handelingen de leraar nu als eerste zou moeten gaan leren en welke handelingen vooralsnog te moeilijk lijken voor de leraar om te leren.

Hoofdstuk 4

Introductie

Evaluatie in scholen kent een variëteit aan methoden. Aan de ene kant is deze variëteit wenselijk omdat het vanuit verschillende invalshoeken informatie kan bieden over het lesgeven van een specifieke leraar. Aan de andere kant blijken de methoden vaak heel verschillende informatie te bieden waarvan onzeker is of deze met elkaar vergeleken of gecombineerd mag worden.

De evaluatiesystematiek die is uitgevoerd in het kader van dit proefschrift maakt gebruik van twee evaluatiemethodes: Lesobservaties en leerlingenvragenlijsten. Sommige wetenschappers hebben gesteld dat het vergelijken van lesobservaties met leerlingenvragenlijsten wat weg heeft van het vergelijken van appels met peren. Ze stellen dat leerlingenvragenlijsten en lesobservaties tot compleet verschillende inzichten over lesgeven kunnen leiden. Anderen stellen het tegenovergestelde, namelijk dat deze twee methodes vergelijkbare informatie bieden en juist gecombineerd moeten worden tot één evaluatie. Zij bieden bewijs voor deze stelling door te laten zien dat evaluaties gebaseerd op een combinatie van lesobservaties en leerlingenvragenlijsten meer betrouwbaar zijn dan evaluaties op basis van enkel lesobservaties.

Uit het onderzoek in Hoofdstuk 2 en 3 is al gebleken dat zowel lesobservatoren als leerlingen een soortgelijke stapsgewijze volgorde aan pedagogische en didactische handelingen toekennen. Dit betekent dat zowel leerlingen als lesobservatoren ‘veilig en stimulerend leerklimaat’ het minst complex vinden (zodat dit het gemakkelijkste als eerste geleerd kan worden), terwijl ze beiden differentiatie en lesgeven in leerstrategieën het meest complex vinden. In dit onderzoek wordt de vergelijkbaarheid van de leerlingenvragenlijst en het lesobservatieformulier verder onderzocht. Om dit na te gaan, worden de items van de vragenlijst en van de lesobservaties samen geanalyseerd alsof ze van hetzelfde instrument komen. Het is dan mogelijk om te onderzoeken of leerlingen en lesobservatoren alle stappen ongeveer even complex vinden, of dat lesobservatoren alle stappen minder/meer complex vinden dan leerlingen. Bijvoorbeeld, zou het kunnen dat lesobservatoren en leerlingen dezelfde stapsgewijze volgorde observeren, maar dat lesobservatoren de stap (2) *efficiënte lesorganisatie* net zo complex vinden als de leerlingen de stap (4) *intensieve en activerende les*? In dit voorbeeld zouden leerlingen hun leraren altijd twee stappen hoger evalueren dan dat de lesobservatoren doen. Dit zou betekenen dat de lesobservatoren en leerlingen geen overeenkomstige interpretatie geven aan de afzonderlijke handelingen en dus dat de resultaten uit de leerlingenvragenlijst niet zomaar kunnen worden opgeteld bij de resultaten van de lesobservaties.

Methode

Voor dit onderzoek is een steekproef gebruikt die bestond uit 269 lesobservaties en 2.876 leerlingenvragenlijsten. De lesobservaties vonden plaats in de klas die ook de leerlingenvragenlijst invulde zodat de observatoren dezelfde lessen observeerden als de leerlingen. Deze data zijn verkregen in de schooljaren 2013-2014 en 2014-2015. Het ICALT-lesobservatieformulier kent 32 items, waarvan 31 items werden meegenomen in de analyse (zie Hoofdstuk 2). De gebruikte versie van de "Mijn Leraar" vragenlijst kent 40 items waarvan – op basis van de resultaten in Hoofdstuk 3 – 28 items zijn meegenomen in de analyse (zie Appendix E). Om te toetsen of de handelingen wanneer gevraagd aan leerlingen en aan lesobservatoren een gelijke stapsgewijze ontwikkeling laten zien is een multiniveau Rasch analyse uitgevoerd.

Resultaten

De resultaten geven aan dat de 31-items van de ICALT en de 28-items van de leerlingenvragenlijst samengenomen kunnen worden tot een eendimensionale stapsgewijze volgorde. Wel blijkt dat een aantal items op de "Mijn Leraar" vragenlijst gerelateerd aan leerstrategieën de ordering wat verstoren. Ook blijkt dat de handelingen gerelateerd aan de stap (of het domein) "differentiatie" wanneer geobserveerd met de ICALT te vaak dezelfde score krijgt. Dus alhoewel de items in de stap differentiatie verschillende handelingen beschrijven worden deze verschillen te weinig opgemerkt door de observatoren. De leerlingen merken meer verschillen op in de handelingen binnen de stap differentiatie.

Een tweede belangrijk resultaat is dat ondanks dat de obervatoren en leerlingen grotendeels overeenstemmen in hun interpretatie van de items, ze wel heel verschillend kunnen denken over of de leraar ook bekwaam is in een pedagogisch of didactische handeling beschreven door dat item. Wanneer een evaluatieprocedure wordt gebruikt waarin de observator slechts één les bezoekt en de uitkomsten van deze ene lesobservatie worden vergeleken met de observaties van leerlingen die alle lessen hebben gezien dan is de correlatie laag ($r = .26$).

Conclusie & Implicaties

De studie in Hoofdstuk 4 biedt aanvullend empirisch bewijs dat lesobservatoren en leerlingen wel degelijk vergelijkbare informatie *kunnen* verschaffen, waaruit geconcludeerd kan worden dat lesobservaties en leerlingenvragenlijsten geen 'appels en peren' zijn. Toch biedt het ook aanvullend bewijs dat niet verwacht kan worden dat één lesobservatie een voldoende representatief beeld geeft van de bekwaamheid van de leraar in lesgeven (zie ook Hoofdstuk 5), waardoor leerlingen – die veel meer lessen hebben gezien – de leraar behoorlijk anders kunnen evalueren.

Met dit onderzoek is een verdere stap gezet in hoe scholen gegevens verkregen met verschillende evaluatiemethoden zouden kunnen combineren. Het blijkt mogelijk om de gegevens verkregen met verschillende methoden op eenzelfde meetschaal te plaatsen. Sommige evaluatie methoden kunnen gemakkelijker en goedkoper op grotere schaal worden ingezet dan anderen. Een mogelijke implicatie van dit onderzoek is dat zulke meer efficiënte methoden zouden kunnen worden gebruikt om een globaal overzicht te krijgen in welke lessen en bij welke leraren de inzet van duurdere (maar ook voor de betreffende docent meer informatieve) vormen van evaluatie waardevol lijkt.

Hoofdstuk 5

Introductie

Wanneer scholen docenten feedback willen geven ofwel hun functioneren willen beoordelen dan is het belangrijk om na te gaan of de evaluaties, waarop deze feedback en besluiten worden gebaseerd, betrouwbaar zijn. Wanneer evaluaties betrouwbaar zijn, dan betekent dit dat de uitkomsten gerepliceerd kunnen worden. Dus wanneer een andere observator een andere les zou hebben geobserveerd, zou de uitkomst voor de betreffende leraar niet anders uitvallen. Wanneer een evaluatie te weinig betrouwbaar is, dan neemt een school een onacceptabel risico dat het een leraar verkeerde feedback teruggeeft of dat zij op oneigenlijke gronden evaluatieve besluiten neemt over de leraar. Vanuit het voorgaande volgt dat er een norm gesteld dient te worden voor wat ‘te weinig’ betrouwbaar is. Het is ook duidelijk dat deze norm anders ligt voor feedback – wat doorgaans geen directe persoonlijke consequenties heeft – dan voor evaluatieve besluiten – wat doorgaans wel directe persoonlijke consequenties hebben. Op basis van deze redenatie zijn er twee criteria gesteld: een betrouwbaarheid van .70 voldoet voor feedback, een betrouwbaarheid van .90 voldoet voor evaluatieve besluiten.

Methode

De studie is gebaseerd op 198 lesobservaties uitgevoerd door 62 lesobservatoren bij 69 leraren werkzaam op 8 verschillende scholen. De lesobservaties zijn gedaan met het

ICALT-observatie instrument dat 32 items telt. Op basis van de resultaten in Hoofdstuk 2 is besloten om 31 items in de analyse mee te nemen.

De analyse van betrouwbaarheid combineert de principes van generaliseerbaarheidstheorie (G-theorie) met de principes van de item response theorie (IRT), tezamen wordt verwezen naar deze analysetechniek als GIRT.

Resultaten

De resultaten tonen aan dat één lesobservatie onvoldoende betrouwbaar is om leraren van feedback te voorzien en zeker onvoldoende betrouwbaar is om te gebruiken voor de onderbouwing van besluiten over leraren in functioneringsgesprekken. Wanneer vier lessen worden bezocht door vier verschillende observatoren is hun gemiddelde uitkomst voldoende betrouwbaar om leraren feedback te geven (hoger dan .70). Na meer dan 10 verschillende lesbezoeken begint de stijging in betrouwbaarheid nihil te worden. De extra lesbezoeken brengen dus weinig nieuwe informatie in. Ook bij 10 lesbezoeken is het gemiddelde niet betrouwbaar genoeg om te gebruiken als input op een functioneringsgesprek (d.w.z. de betrouwbaarheid blijft beneden de .90).

Conclusies en Implicaties

Scholen worden geadviseerd minimaal vier lesbezoeken door vier verschillende observatoren te verzamelen alvorens de leraar feedback te geven. Wanneer scholen dit zonder gebruik te maken van eventueel aanvullende informatie niet doen, nemen ze een onacceptabel groot risico dat ze leraren verkeerde feedback geven op basis waarvan de leraar zich niet kan verbeteren. Als gevolg zal de leraar misschien deelnemen aan trainingen of meedoen in intervisiesessies met verminderde kans op resultaat. Scholen worden verder geadviseerd niet meer dan 10 lesbezoeken per jaar te doen. Het effect van nog meer lesbezoeken is verwaarloosbaar. Het is beter om in plaats van nog meer lesbezoeken andere methodes voor evaluatie in te zetten. Een combinatie van lesobservatie en andere evaluatiemethodes is noodzakelijk om te zorgen dat informatie betrouwbaar genoeg is voor besluiten over het functioneren van de leraar. Wederom geldt dat de school een onacceptabel risico loopt om op verkeerde grond evaluatieve besluiten te nemen over leraren wanneer ze deze baseren op bijvoorbeeld alleen lesbezoeken, ongeacht hoeveel.

Hoofdstuk 6

Introductie

Sommige van de theorieën over de ontwikkeling van leraren berust op de aanname dat er (relevante) individuele verschillen zijn in de ontwikkeling. Hierbij is soms gesuggereerd dat er 'exceptionele' leraren zijn die een talent hebben voor het vak waardoor ze zich het vak op een andere manier eigen maken dan hun minder getalenteerde collega's. Gezien de implicaties die evaluaties kunnen hebben voor individuele leraren is het een morele verplichting na te gaan of er sprake is van individuele verschillen in de ontwikkeling van lesgeven. Als er reden is om aan te nemen dat sommige leraren zich op een exceptionele manier ontwikkelen dan zal de feedback op basis van de gewone ontwikkeling – zoals gespecificeerd in Hoofdstuk 2 en 3 – voor deze leraren niet bijdragen aan hun ontwikkeling. Een leraar die zich exceptioneel ontwikkelt wordt in deze studie gedefinieerd als een leraar die eerst de meer complexere vaardigheden ontwikkelt (bijv. aanleren van leerstrategieën of het differentiëren) voordat de basale vaardigheden zijn ontwikkeld (bijv. een efficiënt klassenmanagement).

Methode

De steekproef bestaat uit dezelfde 198 lesobservaties die ook in Hoofdstuk 5 zijn onderzocht. Eerst werd van alle lessen vastgesteld of er in deze les relatief complexe pedagogische en didactische handelingen werden verricht terwijl in diezelfde les de minder complexe handelingen niet werden geobserveerd. Zulke lessen werden gediagnosticeerd als 'afwijkend'. Vervolgens is nagegaan of lessen waarin sprake was van afwijkend beeld, telkens werden gedoceerd door dezelfde leraren. De analyse – uitgevoerd met de G_{NORMED} person fit statistiek– kan niet worden uitgevoerd voor lesobservaties waarin één item een missende waarde had. Hierdoor konden er van de 198 lesobservaties 141 worden meegenomen. Er is wel nagegaan of verwacht mag worden dat de resultaten wijzigen wanneer de lesobservaties met een missende waarde wel zouden worden meegenomen. Dit lijkt niet het geval.

Resultaten

Van alle 141 lessen werd 15% gediagnosticeerd als afwijkend, maar er kon geen overtuigend bewijs worden gevonden dat afwijkende lessen vaker gedoceerd werden door dezelfde groep 'exceptionele' leraren. De 21 (15%) afwijkende lessen werden gedoceerd door 18 verschillende leraren. Het lijkt er daarom op dat iedere leraar incidenteel een afwijkende les kan hebben.

Conclusies & Implicaties

De resultaten bevestigen eerdere resultaten dat sommige lessen afwijken of ‘exceptioneel’ zijn. Wat er gebeurt in die lessen stemt niet overeen met wat op basis van theorie over de ontwikkeling van leraren wordt voorspeld. Het is op basis van deze steekproef mogelijk dit resultaat verder te analyseren. Hieruit blijkt dat er weinig bewijs is dat zulke afwijkende lessen vaker voorkomen bij een groep ‘exceptionele’ leraren die zich dan anders zouden ontwikkelen dan de reguliere leraar. Daarmee bieden de resultaten weinig bewijs voor de stelling dat er relevante individuele verschillen zijn in de ontwikkeling van leraren waarmee rekening gehouden zou moeten worden in de feedback.

Omdat het erop lijkt dat afwijkende lessen bij iedere leraar kunnen voorkomen is het belangrijk dat scholen leren zulke lessen adequaat te herkennen en verwijderen om zodoende te voorkomen dat deze incidentele afwijkingen de validiteit van de feedback en evaluatieve besluiten verminderen. Deze conclusie bevestigt wat we in hoofdstuk 5 al concludeerde, namelijk dat het nodig is om meerdere lesobservaties bij iedere leraar uit te voeren.

Algemene conclusie proefschrift

De centrale onderzoeksvraag van dit proefschrift is:

Hoe kunnen lesobservaties en leerlingenvragenlijsten op zo’n manier worden ingezet in scholen dat ze leraren op een valide en betrouwbare manier van feedback kunnen voorzien en dat scholen op een valide en betrouwbare manier evaluatieve besluiten over leraren kunnen nemen?

Op basis van de deelstudies kan worden geconcludeerd dat zowel lesobservaties als leerlingenvragenlijsten valide en betrouwbare feedback kunnen geven over het lesgeven van leraren. Zowel de “Mijn Leraar” vragenlijst als de ICALT-lesobservaties tonen eenzelfde stapsgewijze volgorde van het ontwikkelen van pedagogisch en didactische handelingen van leraren. Dit maakt het mogelijk om leraren feedback te geven over in welke vaardigheden ze al bekwaam zijn, welke vaardigheden ze zich nu eerst moeten proberen te bekwamen, en welke vaardigheden vooralsnog te complex zijn om zich in te bekwamen. Op deze manier kan meer specifieke feedback gegeven worden.

Op basis van de deelstudies kan ook worden geconcludeerd dat één enkele lesobservatie onvoldoende is om leraren van feedback te voorzien. Wanneer feedback wordt gegeven op basis van één lesobservatie dan is de kans onacceptabel groot dat een leraar verkeerde feedback ontvangt. Dit kan ertoe leiden dat een leraar bijvoorbeeld wordt verteld bekwaam te zijn in de eerste 2 stappen, zodat zijn leren zich zou moeten richten op stap 3, terwijl de leraar eigenlijk bekwaam is in de eerste 4 stappen. Zulke feedback kan leiden tot het verkeerd besteden van professionaliseringsgelden en ook tot demotivatie bij de betreffende leraar. Betrouwbare feedback vereist minimaal 4 lesobservaties door 4 verschillende observatoren.

Het lijkt ook mogelijk om lesobservaties in te zetten voor functioneringsgesprekken. Echter, dit moet wel op een verantwoorde, dus betrouwbare, wijze gebeuren. Wanneer een schoolbestuur op een betrouwbare manier besluiten wil nemen over haar personeel, dan wordt zij geadviseerd om meer dan 10 lesobservaties uitgevoerd door verschillende observatoren te verzamelen in combinatie met meerdere leerlingenvragenlijsten. Dit lijkt niet mogelijk binnen één schooljaar. We adviseren daarom om deze hoeveelheid te spreiden over enkele jaren. In de tussentijd kunnen de jaarlijks verzamelde lesobservaties wel worden ingezet voor het geven van feedback tijdens besprekingen over de ontwikkeling van een leraar, bijvoorbeeld in de vorm van een voortgangsgesprek.

APPENDIX A

Betrouwbaarheid van rapportcijfers

Preface to the appendix

The initial intention behind this research was to investigate the potential to use teacher-assigned grades to support informed decisions about teaching effectiveness. The intention herein was to use teacher-assigned grades to make a cost-efficient fast first selection of relatively ineffective teachers. More cost-intensive evaluation methods, including student questionnaires and classroom observations, might than be used to follow-up this initial selection in order to further diagnose how to improve effectiveness. This way, the cost-intensive evaluation methods would not have to be applied for every teacher every year. However, the results indicated that teacher-assigned grades are too unreliable to diagnose the effectiveness of teachers (reliability below .30). This is explained by the class average grades which do neither vary much between teachers nor subjects. When consulting some few teachers about these results, they claimed that teacher-assigned grades on occasion are changed at hindsight, after it is known how students performed. They explained this behavior by pressures felt from the school management to achieve acceptable output in terms of students succeeding to pass to the next grade.

Irrespective of whether this explanation is true or false, there currently are no alternative criteria to evaluate the quality of education and as long as these do not exist the current output criterion is *the* sole criterion to evaluate educational effectiveness. Recognizing this situation, the study was redesigned to study the reliability of teacher-assigned grades to evaluate students. The logic behind this choice was that reliability might function as an additional criterion next to the output criterion. The reliability criterion might detect grade adjustment, thereby making this problem transparent. However, the redesigned research does not fit this dissertations' main theme. Because the data were gathered as part of the dissertation research and the study's results are perceived as highly relevant to the field, it eventually was decided to include the work in the appendix.

Samenvatting

In eerder onderzoek wordt gesteld dat docenten subjectieve beoordelaars zijn die zich bij het geven van cijfers niet beperken tot het becijferen van alleen de leerlingvaardigheid, maar cijfers geven voor een mengelmoes (‘hodgepodge’) van eigenschappen: de ‘hodgepodge’hypothese. Ook zouden docenten verschillen in mildheid; de mildheidshypothese. In dit onderzoek worden deze beide hypothesen onderzocht. Voor dit onderzoek zijn bij twee steekproeven proefwerkcijfers verzameld. De eerste steekproef telt 5988 proefwerkcijfers gegeven aan 192 leerlingen gedurende één schooljaar door 64 docenten. De tweede steekproef telt 29462 proefwerkcijfers gegeven aan 306 leerlingen gedurende drie opeenvolgende schooljaren door 52 docenten. Om de beoordelingsbias te onderzoeken werden een G-studie en D-studie uitgevoerd. De resultaten geven geen overtuigend bewijs voor de twee hypotheses. In het algemeen blijkt dat rapportcijfers een redelijk betrouwbaar onderscheid maken tussen minder en meer vaardige leerlingen ($\mathbf{E}\rho^2 \geq$ .70) en een betrouwbare beoordeling geven over de cesuur voldoende-onvoldoende ($\mathbf{\Phi}_\lambda \approx$ .90). Wanneer rapportcijfers op minder dan 8 proefwerken zijn gebaseerd dan is de betrouwbaarheid lager dan het criterium .70. Een aanzienlijk deel van de onbetrouwbaarheid in beoordeling kan worden verklaard door verschillen in de kwaliteit van de proefwerken en niet door mildheid of hodgepodgegedrag in de beoordeling van docenten.

Abstract

In previous research, teachers report that they use a hodgepodge of factors when grading students. This has led researchers to suspect that teacher-assigned grades are inflated by teacher-student interactions; the hodgepodge hypothesis. Teachers also are reported to differ in grading leniency; the leniency hypothesis. In this study these two hypotheses are investigated. Two samples of teachers-assigned grades were gathered. The first sample contained 5,988 grades awarded by 64 teacher to 192 students during one school year. The second sample contained 29,462 teacher-assigned grades awarded to 306 student by 52 teachers during three subsequent school years. Generalizability Theory is used to analyze bias. The results present little evidence to claim that school grades are considerably biased due to hodgepodge grading or teacher leniency. Unreliability in teacher-assigned grades is more due to the tests than due to teachers’ hodgepodge or leniency.

1 Inleiding

In deze studie wordt ingegaan op de beoordeling van leerlingen door docenten. Er is eerder op diverse manieren onderzoek verricht naar het becijferen door docenten (bijv., Bowers, 2009, 2010 2011; Brookhart, 1994, 2004; Marzano, 2002, Randall & Engelhard, 2010). In dit eerdere onderzoek wordt er vanuit gegaan dat docenten subjectieve beoordelaars zijn omdat docenten zich niet beperken tot het becijferen van alleen de leerlingvaardigheid, maar cijfers geven voor een mengelmoes ('hodgepodge') van eigenschappen waaronder de getoonde motivatie, de houding en het gedrag en de groei die de leerling heeft gemaakt (Brookhart, 1994; McMillan, Myran, & Workman, 2002). Ander onderzoek heeft zich gericht op verschillen in mildheid, zodat sommige docenten hun leerlingen becijferen ten aanzien van strengere eisen dan collega-docenten (bijv., Kuhlemeier & Kremers, 2013). Mede door deze veronderstelde subjectiviteit in cijfers is in veel landen een traditie ontstaan om leerlingvaardigheid ook te evalueren op basis van gestandaardiseerde toetsen en in sommige landen, met name Engeland, neemt deze traditie langzaam de evaluatie op basis van cijfers over (Standaert, 2014).

Toch wijst recent onderzoek uit dat juist de 'subjectieve' cijfers grotere voorspellende validiteit hebben voor het toekomstig schoolsucces dan gestandaardiseerde toetsen (Atkinson & Geiser, 2009; Bowers, 2009, 2010; Cliffordson, 2008; Thorsen & Cliffordson, 2008). Ook studies die construct validiteit van schoolcijfers bestuderen geven geen aanleiding om te veronderstellen dat schoolcijfers compleet onbetrouwbaar zijn (bijv., Brennan, Kim, Wenz-Gros, & Sieperstein, 2001; Südkamp, Kaiser, & Möller, 2012). Gerapporteerde correlaties tussen schoolcijfers en prestaties van dezelfde leerlingen op gestandaardiseerde toetsen zijn hoog en liggen doorgaans tussen 0.5 en de 0.6. Ook hieruit kan opgemaakt worden dat schoolcijfers niet een compleet subjectief oordeel zijn welke los staat van andere maten voor schoolsucces.

De literatuur geeft dus een gemengd beeld. Eerder onderzoek benadrukt zowel de onbetrouwbaarheid en subjectiviteit in becijfering, maar tegelijk geeft het indicaties dat schoolcijfers niet compleet subjectief kunnen zijn (i.e., het is niet mogelijk om meermaals hoge correlaties te vinden wanneer één van beide maten compleet onbetrouwbaar en subjectief zou zijn). Opvallend is, echter, dat, ondanks dat we uit voorgaande studies wel verwachtingen kunnen formuleren over de betrouwbaarheid van becijfering en rapportcijfers, geen van de genoemde studies de betrouwbaarheid van schoolcijfers heeft bestudeerd. Als gevolg hiervan weten we nog weinig van de exacte invloed van subjectiviteit op de betrouwbaarheid van gegeven rapportcijfers. Drany & Wilson (2008) merkten hierover vrij recent nog op:

> “in contrast to the situation for trained raters… we know little about the *consistency* [onze cursivering] with which teachers apply the standards contained within a scoring guide to the work their students generate in the classroom” (p. 418).

Dit roept de vraag op hoe groot de invloed van beoordelingsbias is op de betrouwbaarheid van becijfering en de vraag hoe onderzoek naar betrouwbaarheid van schoolcijfers zou kunnen bijdragen aan de kennis over beoordelingsbias van docenten. In het vervolg van de theoretische achtergrond schetsen we een methodiek om beoordelingsbias in proefwerkcijfers te evalueren aan de hand van Generaliseerbaarheidtheorie (Cronbach, Gleser, Rajaratnam, & Nanda, 1972). De hoofdvraag van het artikel is: *In welke mate worden beoordelingen gemaakt op basis van door docenten gegeven proefwerkcijfers vertekend door beoordelingsbias?*

2 Theoretische achtergrond

2.1 Definiëring van beoordeling en beoordelingsbias

We zullen in de loop van de tekst veelvuldig terugkomen op enkele verwante begrippen: observatie, (proefwerk)cijfer, beoordeling, rapportcijfer en besluit. Om onderscheid te maken tussen deze begrippen maken we gebruik van het werk van Hofstee (1999). Hofstee (1999) spreekt van een beoordeling wanneer *mensen, op gezag van anderen, de kwaliteit(en) van iets van iemand vaststellen*. In dit onderzoek zijn het de docenten die op gezag van de school de vaardigheid van leerlingen in een schoolvak vaststellen. Hofstee (1999) beargumenteert dat beoordelingen zouden moeten zijn gebaseerd op meerdere observaties. Zulke observaties vinden in de scholen regelmatig plaats in de vorm van proefwerken en schooloverhoringen. In de tekst gebruiken we de woorden observatie en proefwerk daarom als synoniemen. De proefwerkcijfers die worden toegekend monden uit in een beoordeling in de vorm van een rapportcijfer. Daarom gebruiken we de woorden beoordeling en rapportcijfer ook als synoniemen. Een besluit is gebaseerd op een groter aantal beoordelingen; vaak over verschillende (niet gemakkelijk verenigbare) vaardigheden (Hofstee, 1999). In scholen leiden de rapportcijfers via een van tevoren afgesproken protocol tot besluiten over doubleren of versnellen.

De definitie van Hofstee maakt tegelijk duidelijk dat er sprake is van beoordelingsbias wanneer een rapportcijfer van een leerling is gebaseerd op andere kwaliteiten dan de vaardigheid in een schoolvak. We zijn ons bewust dat deze definitie een specifieke invulling geeft aan het doel van een rapportcijfer – namelijk het zichtbaar maken van de vaardigheid in een schoolvak– en dat dit kan worden gezien als specifieke invulling van het doel van onderwijs. We willen daarom benadrukken dat het niet onze bedoeling is om de impressie te wekken dat vaardigheid in het schoolvak – of in Biesta (2012) 's termen kwalificatie – het enige doel is van onderwijs. Ons uitgangspunt is dat cijfers slechts één eigenschap of doel tegelijk zouden moeten meten en niet een 'hodgepodge' van meerdere doelen. Wanneer onderwijs ook andere doelen, zoals de persoonlijke vorming van de leerling (of in Bieta's (2012) termen socialisatie en subjectivicatie), nastreeft en mee wil wegen in de besluitvorming over doubleren dan staat het de scholen vrij om leerlingen daarin te becijferen. Door de tijd zijn er periodes geweest waarin scholen aparte cijfers gaven voor attitudes en/of motivatie – er zijn zelfs periodes geweest waarin veel waarde werd gehecht aan de cijfers voor attitude en motivatie (Standaert, 2014). De stelling dat cijfers één eigenschap tegelijk zouden moeten meten is overigens vaker ingenomen; onder andere door De Groot en Wijnen (1983), Brookhart (2004) en Randall en Engelhard (2010).

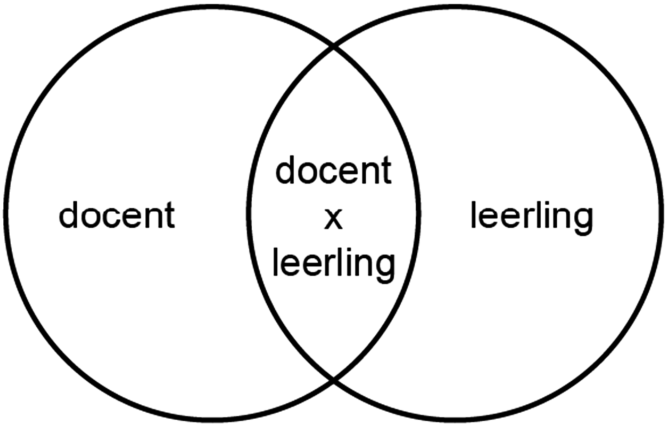
2.2 Beoordelingsbias: Enkele hypotheses en veronderstellingen

In het meeste onderzoek naar beoordelingsbias wordt gebruik gemaakt van zelfrapportage-methodes waarin docenten rapporteren over hun becijferingspraktijk (bijv., Brookhart, 1994, 2004; Cross & Frary, 1999; McMillan, Myran, & Workman, 2002; Randall & Engelhard, 2010). Dit onderzoek bevraagt docenten welke factoren zij meewegen bij het beoordelen van leerlingen. Uit dit onderzoek blijkt dat docenten hun beoordelingen ook van andere factoren laten afhangen dan van de vaardigheid. McMillan et al. (2002) rapporteren dat 39% van de docenten stelt dat zij in de beoordeling van hun leerlingen andere criteria dan de getoonde vaardigheid laten meewegen. In onderzoek van Cross & Frary (1999) stelt 37% van de docenten dat het gedrag en de houding van leerlingen – zoals interesse en nieuwsgierigheid – ook moet meewegen in de beoordeling. Recent zijn deze resultaten herbevestigd door Randall & Engelhard (2010). Op basis van deze resultaten wordt verondersteld dat in de beoordelingen van docenten een mengelmoes van factoren meewegen, waaronder gedrag en attitudes. Deze 'hodgepodge' zou er vervolgens toe leiden dat docenten werk van dezelfde kwaliteit soms toch verschillend becijferen. De resultaten van dit zelfrapportage-onderzoek heeft ruim baan gegeven aan de veronderstelling dat cijfers voor een groot deel worden vertroebeld door docent-leerling-interacties.

De veronderstelling achter de hodgepodgehypothese kan worden weergegeven in een Venn diagram (Figuur 1). De cirkels in Figuur 1 noemen we facetten. Ieder facet geeft een variantie weer. Het facet docent geeft weer dat niet iedere docent tot dezelfde beoordelingen komt. Het facet leerling geeft weer dat niet iedere leerling dezelfde beoordeling krijgt. Het overlappende gedeelte geeft het interactie-effect weer tussen beide facetten. In een situatie waarbij de beoordeling vrij is van bias zou het facet leerling hoge variantie hebben, terwijl de facetten docent en docent × leerling (zeer) kleine variantie hebben.

Figuur 1

Een Venndiagram van het gekruiste design. Wanneer de beoordelingsbias gering is dan zouden de facetten docent en docent × leerling klein zijn.



In de hodgepodgehypothese wordt verondersteld dat een docent een hogere of lagere beoordeling geeft aan één leerling dan aan medeleerlingen, terwijl andere docenten deze ene leerling niet hoger of lager beoordelen. In dit geval zou er hoge variantie zijn in het facet docent-leerling-interactie en is er sprake van bias.

Zoals al opgemerkt, een beperking van het eerdere onderzoek naar de hodgepodgehypothese is dat het zich heeft beperkt tot zelfrapportage-onderzoek. Volgens Muijs (2006) geeft zelfrapportage-onderzoek een imperfecte indicator van het vertoonde gedrag van leraren. Het laat onvoldoende zien hoe de situatie werkelijk is. Het is daarom waardevol om te zoeken naar andere onderzoeksmethoden waarmee de hodgepodgehypothese kan worden onderzocht.

In een andere lijn van onderzoek naar beoordelingsbias wordt verondersteld dat docenten verschillen in strengheid. Dit kan getypeerd worden als de mildheidshypothese (bijv., Drany & Wilson, 2008). In deze hypothese wordt verondersteld dat de variatie in het facet docent (zie Figuur 1) relatief hoog is. Deze hypothese is vooral onderzocht met quasi-experimenteel onderzoek (bijv., Drany & Wilson, 2008; Kuhlemeier & Kremers, 2013; Starch & Elliot, 1914), waarbij hetzelfde werk is beoordeeld door zowel de docent als een tweede beoordelaar. Ook in deze lijn van onderzoek wordt gesteld dat docenten subjectieve beoordelaars zijn, maar de resultaten zijn minder dramatisch. Drany en Wilson (2008) rapporteren bijvoorbeeld dat sommige docenten milder zijn dan andere docenten, maar ook dat er een ‘redelijke mate’ van consistentie bestaat tussen docenten in hun beoordelingen.

De conclusies uit dit eerdere quasi-experimentele onderzoek worden beperkt doordat de onderzoeksprocedure op een aantal punten afwijkt van de praktijksituatie waarin een docent cijfers toekent aan werk van leerlingen, in het bijzonder: (1) in het eerdere onderzoek wordt het werk van leerlingen becijferd door beoordelaars die deze leerling niet kennen, terwijl een docent nooit het werk becijfert van leerlingen die onbekend zijn; (2) in het eerdere onderzoek worden cijfers gegeven die geen consequenties hebben voor de leerling, terwijl cijfers die worden gegeven in de school dit wel hebben; en (3) dit eerdere onderzoek is gebaseerd op de verschillen in becijfering van één proefwerk terwijl op scholen beoordelingen worden gegeven op basis van meerdere proefwerken. Deze beperkingen geven reden tot twijfel in hoeverre de resultaten in dit quasi-experimenteel onderzoek kunnen worden gegeneraliseerd naar de onderwijspraktijk. Op basis van dit eerdere quasi-experimenteel onderzoek kan gesteld worden dat docenten mogelijk verschillen in mildheid bij het eenmalig toekennen van cijfers aan proefwerken, maar het is niet duidelijk hoe docenten verschillen in mildheid in geval van het beoordelen van de leerling op basis van meerdere proefwerken.

Samenvattend kan gesteld worden dat in eerder onderzoek het theoretische model van Figuur 1 in beperkte mate is getoetst. Het empirische onderzoek heeft zich tot nog toe vooral gericht op de mildheidhypothese: de docent is milder naar alle leerlingen dan een collega-docent zou zijn. De hodgepodgehypothese is, voor zover ons bekend, alleen met zelfrapportage onderzocht. Ook worden de conclusies van eerder onderzoek naar de beoordelingsbias beperkt door onderzoeksprocedures die afwijken van de gangbare procedures op school.

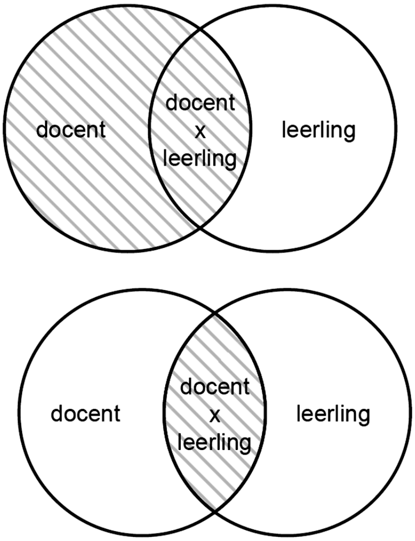
2.3 Vormen van beoordeling; absolute en relatieve beoordeling

Er zijn twee analysemethoden die hierbij gebruikt kunnen worden die gebaseerd zijn op twee manieren om kwaliteiten te beoordelen: (1) op basis van een cesuur of (2) op basis van een rangorde (bijv., De Groot & Wijnen, 1983), ook wel beschreven als respectievelijk criteriumbeoordeling en normatieve beoordeling (bijv., Cliffordson, 2008). In navolging van Cronbach, et al. (1972) refereren we naar de eerste manier van beoordeling met de term 'absolute beoordeling' en de tweede manier van beoordeling met de term 'relatieve beoordeling'. Deze beide manieren van beoordelen leiden tot een andere analyse van de beoordelingsbias (Brennan, 2001; Cronbach, et al., 1972; Shavelson & Webb, 1991). Bij een absolute beoordeling is sprake van beoordelingsbias wanneer docenten geen overeenstemming hebben over het exacte rapportcijfer. Bij relatieve beoordelingen is sprake van beoordelingsbias wanneer docenten geen overeenstemming hebben over de rangorde van leerlingen: van minst naar meest vaardig. De Groot & Wijnen (1983) merken op dat de rapportcijfers in scholen zowel een relatieve beoordelingsfunctie hebben – de rapportcijfers dienen een rangorde weer te geven waarin 'voldoende' leerlingen worden onderscheiden van 'goede' leerlingen – en tegelijk door de cesuur ook absolute beoordelingsfunctie hebben – cijfers onder de 5.5 worden gezien als onvoldoende en cijfers gelijk of groter dan 5.5 als voldoende.

In Figuur 2 is gearceerd wat in een empirische analyse tot de beoordelingsbias wordt gerekend bij een absolute (links) en een relatieve (rechts) beoordeling. Om de beoordelingsbias vast te stellen van een absoluut besluit worden alle facetten die leiden tot afwijkingen in de beoordeling meegewogen. In het linker design worden daarom én het facet docent en het interactie facet docent × leerling beschouwd als bron van bias. Om bias in relatieve besluiten vast te stellen worden alleen de facetten meegewogen die leiden tot wisselingen in de rangorde van leerlingen. In het rechter design wordt daarom alleen het interactie facet beschouwd als bias.

Figuur 2

Een Venndiagram weergave waarin voor absolute en relatieve beoordelingen is gearceerd welke facetten worden gerekend tot beoordelingsbias.



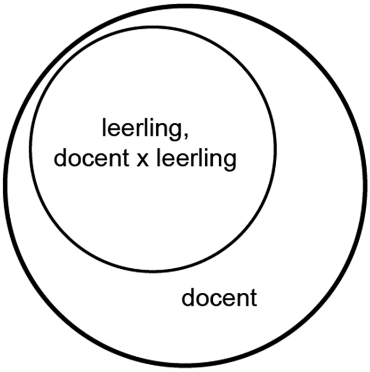
2.4 De schoolpraktijk; van theoretische analyse naar de praktische mogelijkheden

Het is moeilijk om een quasi-experimenteel design op te zetten waarin zowel de docent als een tweede (en derde) beoordelaar bekend zijn met de leerling en waarin deze docenten bij dezelfde leerlingen meerdere observaties uitvoeren (c.q. meerdere proefwerken geven en becijferen) om tot een beoordeling te komen. Toch zou een dergelijk design nodig zijn om de beoordelingsbias te meten. Een mogelijk alternatief om dit design te operationaliseren zou gevonden kunnen worden in het gebruik van bestaande proefwerkcijfers. In scholen is veel data, in de vorm van proefwerkcijfers, aanwezig waarmee beoordelingsbias zou kunnen worden onderzocht.

Deze analyse met proefwerkcijfers is niet zonder problemen. Het belangrijkste obstakel is dat het op scholen gemeengoed is dat proefwerkcijfers worden toegekend aan één klas door één docent zonder dat andere collega’s dit werk (steekproefsgewijs) voor een tweede maal becijferen. We weten daarom hoe één docent de proefwerken becijfert, maar missen informatie over hoe andere docenten diezelfde proefwerken zouden becijferen. Deze voor scholen gangbare procedure leidt tot een genest design (zie Figuur 3) in plaats van het meer volledige en in de theoretische analyse getoonde gekruiste design (Figuur 1).

Figuur 3

Een Venndiagram van het geneste design. Kenmerkend is dat het facet leerling binnen het facet docent zit. Als gevolg is het facet docent × leerling interactie *confounded* met het facet leerling. Dit is aangegeven door beide te noemen in dezelfde cirkel met daartussen een komma.



Het belangrijkste kenmerk van het geneste design is dat de interactie tussen docent en leerling *confounded* is met het facet leerling. De *confound* is aangegeven in de Figuur 3 door beide facetten leerling en docent × leerling, te benoemen in dezelfde cirkel met daartussen een komma. Deze *confound* houdt in dat de variantie in cijfers tussen leerlingen wordt opgeteld bij de variantie in cijfers die ontstaat door docent × leerling interacties, zodat het facet leerling nu twee verklaringen heeft: (1) er is verschil in vaardigheid tussen leerlingen en (2) docenten beoordelen leerlingen van gelijke vaardigheid verschillend. Door deze dubbele verklaring is het niet goed mogelijk om de hodgepodgehypothese te toetsen met het geneste design. Wel is het mogelijk om de mildheidshypothese te toetsen.

Er is een aantal mogelijkheden om met de cijfers beschikbaar op de scholen toch tot een gekruiste analyse te komen, maar deze mogelijkheden zijn gebaseerd op assumpties. De eerste mogelijkheid is om cijfers van dezelfde leerlingen te analyseren bij meerdere vakken. In dit geval hebben meerdere docenten de vaardigheid van de leerling beoordeeld en wanneer één docent tot een andere beoordeling komt dan de collega's zou dit duiden op bias. Lastig is dat de verschillen in beoordeling ook het gevolg kunnen zijn van verschillen in vaardigheid tussen de schoolvakken. Deze eerste mogelijkheid heeft dus de assumptie dat leerlingen niet noemenswaardig verschillen in vaardigheid tussen vakken: een leerling is goed in school of niet. Deze assumptie wordt niet ondersteund door eerder empirisch onderzoek (bijv., Bowers, 2011; Korobko, Glas, Bosker & Luyten, 2008; Thorsen & Cliffordson, 2008) waaruit blijkt dat verschillen tussen beoordelingen voor een belangrijk deel kunnen worden verklaard doordat leerlingen verschillen in hun vaardigheid tussen vakken.

Een andere mogelijkheid is om schoolcijfers van diverse schooljaren op te vragen. Leerlingen kunnen wisselen van docent bij opeenvolgende schooljaren. In deze methode wordt gebruik gemaakt van een specifieke eigenschap van ons becijferingsysteem, namelijk dat cijfers voor ieder proefwerk en ieder schooljaar opnieuw worden geijkt. De onderliggende assumptie is dat de vaardigheid van leerlingen in dezelfde mate toeneemt over de schooljaren. Het gevolg is dat wanneer een leerling een rapportcijfer 6.0 zou halen in het eerste schooljaar, dezelfde leerling wederom een rapportcijfer 6.0 zou halen in het tweede schooljaar. Natuurlijk is de leerling 'verbeterd', maar het werk in het tweede schooljaar is ook 'moeilijker' waardoor de 'verbetering' toch weer uitmondt in een 6.0. Ook deze assumptie is discutabel – er wordt gebruik gemaakt van een eigenschap die beschouwd kan worden als een zwakte van het systeem – maar toch is er vooralsnog geen empirisch bewijs dat aantoont dat deze aanname niet klopt. In dit onderzoek zal daarom worden verkend of deze methode bruikbaar kan zijn om meer inzicht te krijgen in de hodgepodgehypothese.

2.5 Alternatieve interpretaties van de facetten

Tot nog toe is er gesproken over de facetten docent en docent × leerling interactie als indicators van beoordelingsbias. Dit is gangbaar in literatuur rondom beoordeling en becijfering. Toch zijn er alternatieve interpretaties denkbaar. We gaan hier kort in op de belangrijkste van deze alternatieve interpretaties en bespreken in de laatste alinea de gevolgen die deze alternatieve interpretaties hebben voor deze studie.

Het facet docent beschrijft alle verschillen in rapportcijfers tussen docenten. Behalve door beoordelingsbias kunnen zulke verschillen ook ontstaan omdat één docent een hogere kwaliteit van instructie heeft dan een andere docent. Een alternatieve interpretatie voor het facet docent is dus dat deze de verschillen in kwaliteit van instructie beschrijft (bijv., Hattie, 2009). Ook zouden we deze verschillen kunnen interpreteren als gevolg van een verschil in instroom. Wanneer klassen aan het begin van het schooljaar verschillen in niveau, dan zullen deze verschillen waarschijnlijk leiden tot lagere of hogere rapportcijfers aan het einde van het schooljaar (bijv., Wright, Horn, & Sanders, 1997).

Ook het docent × leerling interactie facet, dat betrekking heeft op dat één docent één leerling een rapportcijfer toekent dat een andere docent niet zou toekennen aan die ene leerling, kan zowel wijzen op bias als op een legitiem verschil. Zo kunnen legitieme verschillen ontstaan door succesvolle differentiatie in de instructie. Bij differentiatie in instructie besteedt een docent meer tijd aan een leerling die dat nodig heeft. Wanneer twee docenten verschillen in hun keuze welke leerling meer tijd nodig heeft, of in het geval dat één docent wel succes heeft met de differentiatie maar de andere docent niet, zal dit leiden tot een legitiem verschil in de rapportcijfers.

We stellen vast dat de facetten docent en docent × leerling interactie dus niet alleen beoordelingsbias voorstellen, maar ook informatie kunnen bevatten over legitieme verschillen zoals verschil in kwaliteit van instructie, verschillen in instroom en verschil in differentiatie. Deze legitieme verschillen worden in deze studie meegewogen als bias. Het gevolg is dat in deze studie de mate van bias wordt overschat.

Op basis van bovenstaande richten we ons bij het beantwoorden van de hoofdvraag: *In welke mate worden beoordelingen op basis van door docenten gegeven proefwerkcijfers vertekend door beoordelingsbias?* op de volgende vier deelvragen:

1. In welke mate verschillen docenten in mildheid van beoordelen?
2. In welke mate wordt beoordeling door docenten vertekend door hodgepodge?
3. Hoe betrouwbaar zijn rapportcijfers voor relatieve beoordeling?
4. Hoe betrouwbaar zijn rapportcijfers voor absolute beoordeling onvoldoende - voldoende?

3 Methode

In dit onderzoek wordt ingegaan op de beoordelingsbias van docenten door een analyse van de betrouwbaarheid in schoolcijfers. Als methode gebruiken we Generaliseerbaarheids-theorie (Brennan, 2001; Shavelson & Webb, 1991). In deze methode wordt een G-studie uitgevoerd gevolgd door een D-studie. In een G-studie wordt de grootte van de facetten geanalyseerd. Deze grootte wordt uitgedrukt in een percentage variantie die door dit facet kan worden verklaard. De aandacht in de analyse gaat uit naar de grootte van de beoordelingsbias; de facetten docent en docent × leerling interactie. In de D-studie wordt geanalyseerd wat de gevolgen zijn van de beoordelingsbias op de betrouwbaarheid van de rapportcijfers. Het artikel bestaat uit twee deelstudies. In de eerste deelstudie analyseren we de beoordelingsbias in het geneste design. In het geneste design kan alleen de mildheidshypothese worden getoetst en kan dus niet een compleet antwoord geven op onze hoofdvraag. Deze eerste studie heeft twee functies; (1) het verkennen van beoordelingsbias vanuit een herkenbaar startpunt en (2) de resultaten uit de eerste studie kunnen in de tweede studie worden gevalideerd wat de resultaten van beide deelstudies versterkt. In de tweede studie analyseren we beoordelingsbias in een gekruist design. In deze tweede studie wordt de hodgepodgehypothese getoetst.

3.1 Studie 1

3.1.1 Steekproef en procedure

Voor dit eerste onderzoek zijn 19461 cijfers verzameld. De steekproef bestaat uit alle cijfers die zijn gegeven aan 391 leerlingen in het voortgezet onderwijs (VO) gedurende één schooljaar. De tijdstippen waarop werd becijferd verschilde per klas. Deze 19461 cijfers zijn gevrijwaard van herkansingen. Verder zijn ook de kerst-, paas-, en zomerrapport gemiddeldes uit de cijferbestanden verwijderd.

Het aantal cijfers per leraar schommelde van 29 cijfers in één schooljaar (vak Engels) tot 5 cijfers in één schooljaar (vak geschiedenis). Ook binnen vakken is er behoorlijke variatie in het aantal cijfers die docenten geven: collega-docenten Engels op andere scholen gaven bijvoorbeeld soms 11 cijfers. Dit toont de behoorlijke disbalans in de data en deze disbalans zou de resultaten – met name de grootte van de facetten – kunnen vertekenen (Brennan, 2001). Om dit te ondervangen is er gekozen om per docent acht cijfers willekeurig te selecteren. In het geval dat een leraar minder dan acht cijfers had gegeven gedurende het schooljaar werden alle cijfers geselecteerd. Na deze selectie bleven er 12190 cijfers over.

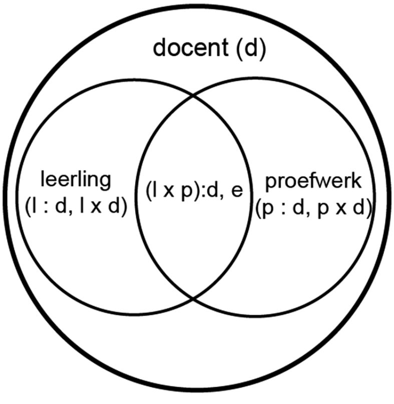
De cijfers zijn gegeven door 64 docenten die lesgaven aan 16 verschillende klassen verspreid over 7 scholen. Voor iedere klas zijn de cijfers voor geschiedenis, wiskunde, Nederlands en Engels verzameld. De leerlingen zaten in de onderbouw (11 – 15 jaar) van het voortgezet middelbaar beroepsonderwijs (VMBO), het hoger algemeen vormend onderwijs (HAVO) of het wetenschappelijk voortgezet onderwijs (VWO). De klassengrootte varieerde van 15 tot 29 met een gemiddelde van 25 waarbij er een sterk verband is tussen de klassengrootte en de onderwijssoort: in het VMBO telden alle klassen minder dan 25 leerlingen. Opnieuw is er sprake van een disbalans, ditmaal door verschil in klassengrootte. Gezien de vraag of er variatie is tussen docenten, en gezien de variatie tussen klassen niet goed kan worden onderscheiden van variatie tussen docenten, is besloten om deze disbalans te reduceren door van iedere klas 12 leerlingen willekeurig te selecteren. Van de oorspronkelijke 391 leerlingen werden er 192 geselecteerd. Het aantal cijfers daalde hierdoor van 12190 tot 5988.

3.1.2 Design

Het design heeft een geneste structuur (zie Figuur 4). De observaties bij een leerling zijn genest in docent (l : d). De ":" betekent 'genest in'. Naast dat leerlingen genest zijn in docenten heeft het design ook een facet proefwerk (p). Deze proefwerken zijn ingevoerd in de data als een *long form* (de Boeck, et al. 2011). Bij een *long form* worden alle proefwerkcijfers onder elkaar gezet in één kolom. Proefwerken worden beschouwd als gekruist met het facet leerling: alle leerlingen bij een docent hebben dezelfde proefwerken gemaakt. Ook worden proefwerken beschouwd als genest in docent: iedere docent geeft zijn of haar leerlingen andere proefwerken. Het interactie facet proefwerk × docent (p × d) is *confounded* met het facet proefwerk. Het interactie facet docent × leerling is *confounded* met het facet leerling. Het interactie facet leerling × proefwerk (l × p) is *confounded* met het residu (*e*).

Figuur 4

Een Venndiagram weergave van het geneste design van de eerste studie (l × p): d



3.1.3 Data analyse

In de eerste studie concentreren we ons op de mildheidshypothese. Analyses voor de G-studie zijn uitgevoerd in R met het package lme4 (Bates, Maechler, Bolker, & Walker, 2014). Om met lme4 een G-studie uit te voeren werd een random effects model gespecificeerd. De D-studie is gebaseerd op de output uit de G-studie. De betrouwbaarheidscoëfficiënten zijn berekend op basis van formules zoals in Brennan (2001) en Shavelson en Webb (1991). Voor de relatieve betrouwbaarheid geldt:

$$\boldsymbol{E}(\rho^2) = \frac{\sigma^2_{(\mathrm{l:d})}}{\sigma^2_{(\mathrm{l:d})} + \frac{\sigma^2_{(\mathrm{l\times p):d,e})}}{n_\mathrm{p}}} \; (1),$$

We hebben de absolute betrouwbaarheid uitgerekend relatief aan de cesuur van 5.5. Deze keuze is gemaakt omdat absolute besluiten meestal het meer dichotome karakter hebben van onvoldoende of voldoende. Voor de absolute betrouwbaarheid met afkappunt (λ) hebben we de formule gehanteerd:

$$\boldsymbol{\Phi}_\lambda = \frac{\sigma^2_{(\mathrm{l:d})} + ((\mu_{\mathrm{vak}} - \lambda)^2 - (\sigma^2_\mathrm{T}))}{(\sigma^2_{(\mathrm{l:d})} + ((\mu_{\mathrm{vak}} - \lambda)^2 - (\sigma^2_\mathrm{T}))) + \frac{\sigma^2_{(\mathrm{p:d})}}{n_\mathrm{p}} + \frac{\sigma^2_{(\mathrm{l\times p):d,e})}}{n_\mathrm{p}}} \; (2),$$

Waarin λ = criterium voldoende-onvoldoende = 5.5; μ_{vak}= het gemiddelde van het schoolvak; $\sigma^2_{\mathrm{leerlingen\ (l:d)}}$= de variantie tussen leerlingen; $\sigma^2_{\mathrm{proefwerk\ (p:d)}}$ = de variantie tussen proefwerkcijfers; $\sigma^2_{\mathrm{residu\ (l\times p):d,e)}}$ = alle overige variantie; σ^2_T = de optelsom van de drie voorgenoemde facetten gedeeld door hun aantal levels. De subscripts duiden de nesting en kruising van de diverse facetten aan.

3.2 Studie 2

3.2.1 Methode

In het geneste design kan niet worden vastgesteld wat de mate van beoordelingsbias is die ontstaat doordat docenten gelijk werk van twee leerlingen toch ongelijk beoordelen, bijvoorbeeld vanwege verschillen in de getoonde motivatie, persoonlijkheid of getoonde groei. Deze vormen van bias komen juist in zelfrapportage-onderzoek onder docenten prominent naar voren (bijv., Brookhart, 1994; McMillan, et al., 2002). In deze tweede studie verkennen we een onderzoeksopzet waarin het wel mogelijk is de invloed van zulk hodgepodgegedrag op de beoordeling te onderzoeken.

3.2.2 Steekproef en procedure

De steekproef voor de tweede studie bestond uit 56663 cijfers afkomstig van 2 scholen. Deze cijfers zijn gegeven aan proefwerken van alle proefwerksoorten, zoals luistertoetsen, schriftelijke overhoringen, presentaties, praktische opdrachten die werden gegeven aan 424 leerlingen gedurende drie opeenvolgende schooljaren in de onderbouw van de HAVO en VWO. Van deze 424 leerlingen waren van 306 leerlingen voor minimaal twee van de drie opeenvolgende jaren cijfers beschikbaar. De overige 118 (27.7%) leerlingen verhuisden naar een andere schoolsoort of doubleerden. De 118 verhuizers en doubleurs waren ongelijk verdeeld over de 3 opeenvolgende schooljaren; de meeste leerlingen verhuisden tussen het tweede en het derde schooljaar (18.7%). Hierdoor waren van deze leerlingen alleen cijfers beschikbaar in hun derde schooljaar. Enkele leerlingen verhuisden na het brugjaar (2.8%) waarna ze ofwel doubleerden ofwel nogmaals verhuisden, zodat van hen alleen cijfers beschikbaar waren in het tweede schooljaar. De overige 6.5 % verhuisde of doubleerde na het brugjaar. De 118 verhuizers waren ongelijk verdeeld over de klassen; waarbij de meeste klassen 1 tot 5 verhuizers of doubleurs telden. In het derde schooljaar werden, echter, soms bijna complete klassen geformeerd met verhuizers (20 leerlingen). Ook waren de rapportcijfers van deze 118 verhuizers of doubleurs meer homogeen (*SD* = .66) dan de rapportcijfers van de overige leerlingen (*SD* = .74). Uiteindelijk is besloten om de 118 verhuizers in de analyse te laten. We bespreken de gevolgen hiervan voor de interpretatie van de resultaten in de discussiesectie beperkingen.

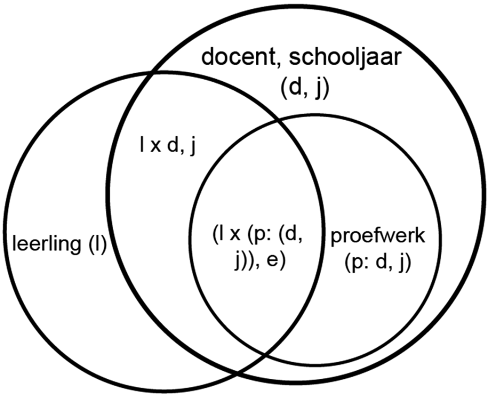
Deze 56663 cijfers zijn gegeven door 52 docenten, waarvan 10 docenten geschiedenis, 12 docenten wiskunde, 16 docenten Nederlands en 15 docenten Engels. Ook in deze steekproef gaven de talen in het algemeen meer cijfers dan wiskunde en geschiedenis en om de disbalans te verkleinen werd besloten om willekeurig 8 cijfers per vak te selecteren. Dit resulteerde in een steekproef van 29462 cijfers. De klassengrootte varieerde van 22 tot 31.

3.2.3 Design

In het tweede design (zie Figuur 5) zijn proefwerkcijfers over meerdere schooljaren verzameld. Het verschil met het 'geneste' design in de eerste studie is dat dit design een apart facet heeft voor de docent × leerling ((d, j) × l) interactie.

Figuur 5

Een Venndiagram weergave van het gekruiste design van de tweede studie: *l* × *(p : (d, j))*.



Omdat leerlingen wisselen van docent in opeenvolgende schooljaren kan dit design nagaan of één leerling door een docent hoger wordt beoordeeld dan dat deze leerling wordt beoordeeld door andere docenten in de vakgroep. Omdat iedere nieuwe docent altijd samengaat met een nieuw schooljaar is het facet docent in deze studie *confounded* met het facet schooljaar (j). Deze *confound* houdt in dat de variantie in cijfers tussen schooljaren wordt opgeteld bij de variantie in cijfers tussen docenten, zodat de variantie in het facet docent twee verklaringen heeft: (1) er is verschil tussen docenten en (2) er is verschil tussen de schooljaren. De assumptie is dat het facet docent vooral verschillen tussen docenten weergeeft, terwijl de schooljaren nauwelijks bijdragen aan de verschillen in het facet docent. Om na te gaan of deze assumptie verdedigbaar is werden de leerlingen geselecteerd die in opeenvolgende jaren dezelfde docenten hadden. In deze groep geeft het facet docent alleen verschillen weer tussen schooljaren en uitgaande dat de assumptie klopt zouden de correlaties de 1.00 moeten naderen. Omdat de leerlingen niet in beide jaren dezelfde proefwerken hebben gemaakt werd de correlatie berekend met de gemiddelde rapportcijfers. De Pearson correlaties tussen de rapportcijfers voor deze leerlingen waren: $r = .70$ (jaar 1 en jaar 2) $r = .67$ (jaar 2 en jaar 3) en $r = .88$ (jaar 1 en jaar 3). Deze correlaties naderen inderdaad de 1.00. Ook waren de correlaties beduidend hoger dan de Pearson correlaties voor de leerlingen die wel wisselden van docent in opeenvolgende jaren; *r* = .66, *r* = .37 en *r* = .55 respectievelijk. De verschillen in rapportcijfers tussen opeenvolgende schooljaren lijken dus inderdaad vooral te informeren over de wisselingen in docent. Net als in het geneste design zijn in dit design de andere interacties (docent × proefwerk (d × p) en leerling × proefwerk (l × p)) *confounded*.

3.2.4 Data analyse

De data analyse met het tweede design richt zich op hodgepodgehypothese. De analyse voor de G-studie is wederom gedaan met het lme4 package van R (Bates, et al., 2014). De D-studie is gebaseerd op de output van de G-studie. In de D-studie richten we ons op de relatieve betrouwbaarheid. De gebruikte formule is:

$$\boldsymbol{E}(\rho^2) = \frac{\sigma^2_{(\mathrm{l})}}{\sigma^2_{(\mathrm{l})} + \frac{\sigma^2_{(\mathrm{d,j} \times \mathrm{l})}}{n_{\mathrm{d}}} + \frac{\sigma^2_{\mathrm{l} \times (\mathrm{p:(d,j)})}}{n_{\mathrm{p}} n_{\mathrm{d}}}} \ (3),$$

4 Resultaten Studie 1

4.1.1 Beoordelingsbias door mildheid

In de Tabel 1 hieronder worden de resultaten van de G-studie met drie facetten (docent, leerling, proefwerk) gepresenteerd. Alhoewel Tabel 1 meer informatie geeft beperken we ons tot het bespreken in hoeverre de resultaten indicaties geven van beoordelingsbias. Een indicatie van de mate waarin verschillen in mildheid tussen docenten de beoordeling vertroebelen wordt gegeven door de grootte van het facet docent (d). De resultaten geven weer dat verschillen in mildheid tussen docenten klein tot verwaarloosbaar is (d.w.z. variërend van 0 - 7%). Omdat het facet docent (d) ook een indicatie geeft van legitieme verschillen in de kwaliteit van lesgeven tussen docenten en de verschillen in instroom, kan geconstateerd worden dat de beoordelingsbias door docenten geen substantiële invloed heeft op de hoogte van de rapportcijfers van leerlingen.

Tabel 1

Resultaten G-studie met het geneste design (l × p): d

Facet	geschiedenis		wiskunde		Nederlands		Engels	
	%	**CI(%)**	**%**	**CI(%)**	**%**	**CI(%)**	**%**	**CI(%)**
docent (d)	.07	.036 - .160	.02	.010 - .048	.00	.003 - .013	.00	.000 - .000
proefwerk (p)	.22	.169 - .283	.21	.169 - .277	.25	.182 - .299	.20	.183 - .274
leerling (l : d)	.27	.221 - .331	.29	.240 - .359	.19	.152 - .227	.22	.161 - .266
residu (l × p) : d, *e*	.45	.417 - .487	.47	.443 - .512	.63	.543 - .626	.57	.535 - .616

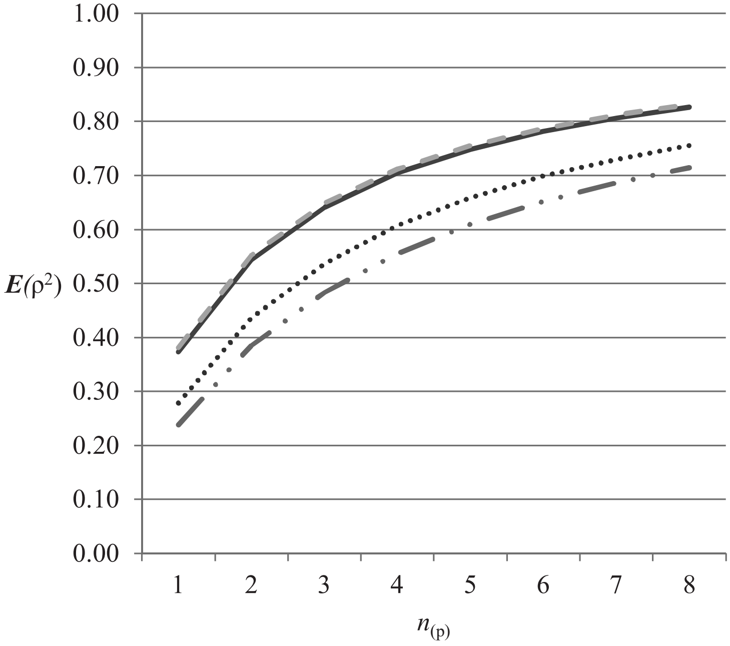
4.1.2 Relatieve betrouwbaarheid van de jaarlijkse rapportcijfers

De resultaten geven ook een indicatie van de mate waarin cijfers worden verstoord door alle vormen van bias inclusief beoordelingsbias. Dit kan worden geanalyseerd met een D-studie. Een D-studie geeft een indicatie van de betrouwbaarheid van de rapportcijfers. De berekende betrouwbaarheid is een gemiddelde en deze kan fluctueren per steekproef (Brennan, 2001; Cronbach, et al., 1972). Daarom wordt in G-theorie gesproken over de verwachte betrouwbaarheid. We presenteren twee D-studies. De eerste voor de betrouwbaarheid van de relatieve beoordeling ($\boldsymbol{E}\rho^2$) en de tweede voor de betrouwbaarheid van de absolute beoordeling voor de cesuur voldoende-onvoldoende ($\boldsymbol{\Phi}_\lambda$). In Figuur 6 wordt de verwachte betrouwbaarheid voor relatieve beoordelingen weergegeven voor ieder vak apart. Hierin is de verwachte betrouwbaarheid voor geschiedenis weergegeven met de ononderbroken lijn, voor wiskunde met de streepjeslijn, voor Nederlands met de stippel-streepjeslijn, en voor Engels met de stippellijn.

Uit Figuur 6 blijkt dat de bias in rapportcijfers afneemt naarmate er meer cijfers worden gegeven, maar ook dat de betrouwbaarheid van beoordelingen bij wiskunde en geschiedenis een hoger plafond heeft dan bij de talen. Dit zou het gevolg kunnen zijn van de grotere variatie in (deel)vaardigheden die in de talen worden getoetst: mondelinge taalvaardigheid, schrijfvaardigheid, leesvaardigheid, literatuur, grammatica en spelling.

Figuur 6

Grafiek van de D-studie gebaseerd op eerste geneste design (l × p): d. Op de y-as de verwachte relatieve betrouwbaarheid (***E***ρ^2). Op de x-as het aantal proefwerken. De ononderbroken lijn = geschiedenis, de streepjeslijn = wiskunde, de stippel-streepjeslijn = Nederlands, en de zwarte stippellijn = Engels



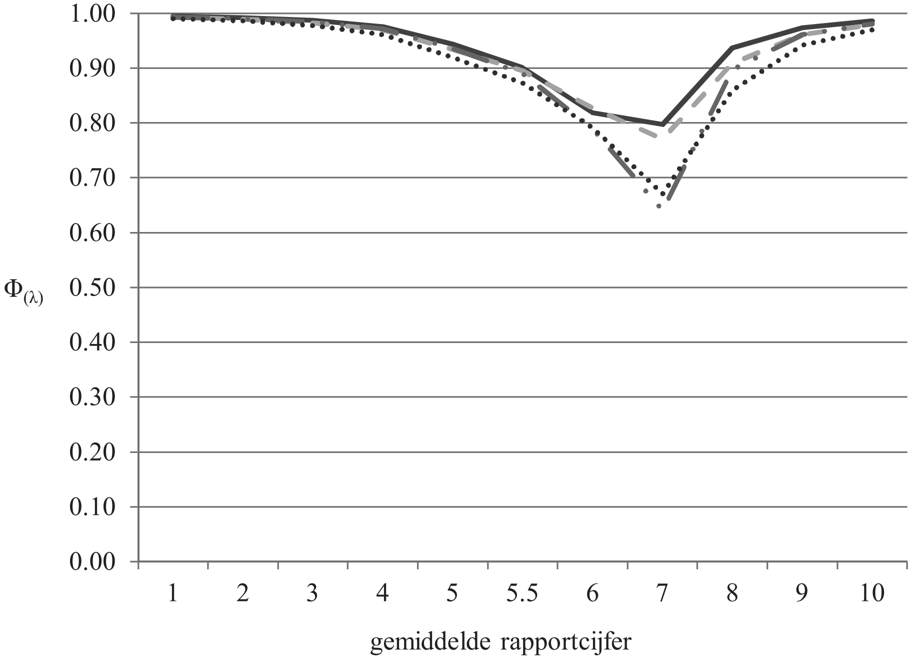
4.1.3 Absolute betrouwbaarheid van de jaarlijkse rapportcijfers

Bij absolute besluiten gaat het erom hoe zeker we zijn dat een rapportcijfer lager of hoger is dan een gegeven criterium; in dit geval 5.5. We doen hiervoor opnieuw een D-studie (zie Figuur 7).

De resultaten laten zien dat docenten op een betrouwbare manier de excellente (> 9.0) en de ‘onvoldoende’ leerlingen (< 5.5) beoordelen. De laagste betrouwbaarheid is bij een rapportcijfer 7.0. Leerlingen die een rapportcijfer 7.0 krijgen toegekend hebben relatief veel proefwerkcijfers die hoger of lager dan een 7.0 waren. Daarom neemt de onzekerheid toe of een leerling echt een rapportgemiddelde 7.0 zou moeten krijgen of eerder een 6.0 of een 8.0. Deze resultaten gaan er vanuit dat er acht proefwerken gedurende het jaar zijn gegeven.

Figuur 7

De grafiek van de D-studie naar de betrouwbaarheid van absolute beoordelingen. Op de y-as de absolute betrouwbaarheid voor een cesuur besluit voldoende-onvoldoende ($\Phi_{(\lambda)}$). Op de x-as het gemiddelde rapportcijfer op basis van 8 cijfers. De ononderbroken lijn = geschiedenis, de streepjeslijn = wiskunde, de stippel-streepjeslijn = Nederlands, en de stippellijn = Engels



Uit de Figuur 7 blijkt dus dat er consistentie is over de cesuur onvoldoende - voldoende. Wanneer er acht (of meer) proefwerken gedurende het schooljaar worden gegeven is de verwachte betrouwbaarheid van een beoordeling over de cesuur voldoende-onvoldoende voor geschiedenis: $\mathbf{\Phi_\lambda}$ = .90, voor wiskunde: $\mathbf{\Phi_\lambda}$ = .90, voor Nederlands: $\mathbf{\Phi_\lambda}$ = .89 en voor Engels: $\mathbf{\Phi_\lambda}$ = .87. De cijfers van de meeste leerlingen liggen dus consistent boven of consistent onder het criterium van 5.5, waardoor docenten voor de meeste leerlingen met behoorlijke zekerheid tot een beoordeling kunnen komen over doubleren.

4.2 Resultaten studie 2

4.2.1 Beoordelingsbias door 'hodgepodge' beoordeling

De resultaten in Tabel 2 geven een eerste inzicht in de grootte van de beoordelingsbias door de docent × leerling interactie in schoolcijfers. Deze eerste verkenning bij twee scholen en 52 docenten wijst erop dat deze interactie slechts een klein deel van de variantie in schoolcijfers kan verklaren (3 - 7%). Ervan uitgaand dat de variatie in het facet docent × leerling ook legitieme verschillen door differentiatie in instructie weergeeft, blijft er slechts ruimte voor een zeer klein percentage gevallen waarin docenten individuele leerlingen hoger beoordelen op basis van motivatie, persoonlijkheid of andere kenmerken anders dan de leerlingvaardigheid. Als er al sprake is van hodgepodge-beoordeling dan lijkt deze dus inconsistent over proefwerken.

Tabel 2

Resultaten G-studie van het tweede design: l × (p : (d, j))

	geschiedenis		**wiskunde**		**Nederlands**		**Engels**	
Facet	**%**	**CI(%)**	**%**	**CI(%)**	**%**	**CI(%)**	**%**	**CI(%)**
docent (d, j)	.07	.037 - .096	.03	.020 - .053	.09	.059 - .153	.03	.021 - .055
proefwerk (p)|	.19	.162 - .227	.14	.119 - .165	.16	.136 - .189	.16	.138 - .192
leerling (l)	.23	.205 - .269	.30	.268 - .351	.17	.149 - .197	.26	.229 - .304
docent × leerling (d, j × l)	.07	.063 - .078	.07	.062 - .076	.04	.034 - .042	.03	.029 - .036
Residu (l × (p : d, j), *e*)	.45	.437 - .468	.45	.438 - .467	.55	.529 - .565	.51	.497 - .530

Daarmee kan geconstateerd worden dat de mate van beoordelingsbias die ontstaat doordat leraren gelijkwaardig werk toch anders becijferen – een praktijk die naar voren komt uit zelfrapportage-onderzoek onder docenten – in deze steekproef niet zo sterk naar voren komt. Misschien dat één derde van de docenten dit doet, maar ze doen dit dan wel met maar enkele leerlingen en de grootte van deze bias is in de orde van tienden op het uiteindelijke rapportcijfer.

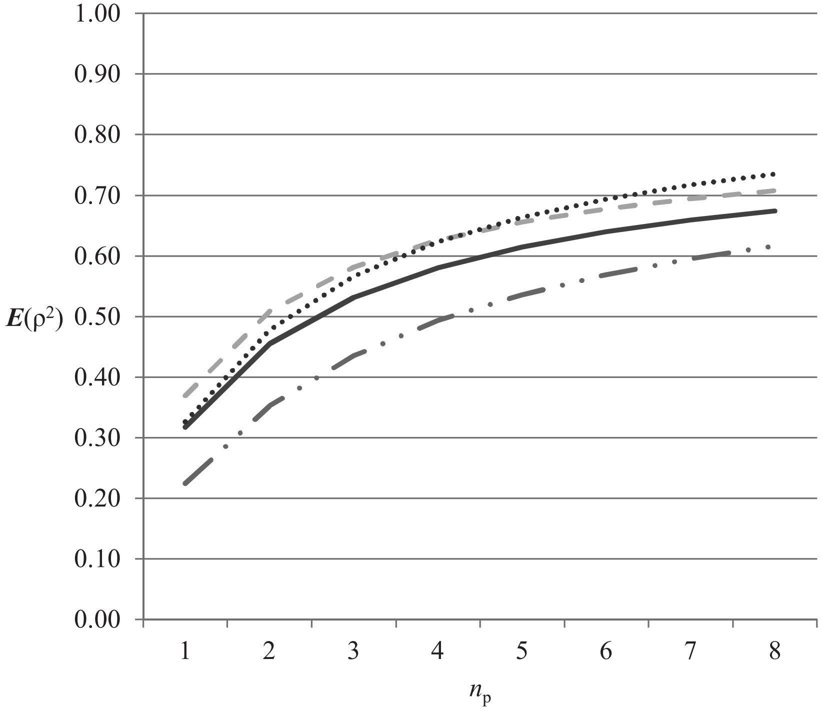
Dit design en deze steekproef bevestigen ook de eerdere resultaten over de mildheidshypothese. Ook hier zijn de percentages klein tot verwaarloosbaar (3 – 8%) en kan geconstateerd worden dat deze kleine verschillen in beoordeling tussen docenten niet alleen het gevolg zijn van bias, maar waarschijnlijk ook van legitieme verschillen. Er lijkt ook hier weinig reden om te argumenteren dat schoolcijfers in hoge mate bias vertonen.

Als laatste stellen we vast dat in beide designs de meeste variantie niet kan worden toegeschreven aan de meegewogen facetten. Het percentage residuele variantie schommelt in beide designs tussen de .45 en de .63. Hieruit kunnen we concluderen dat één enkel proefwerkcijfer maar zeer beperkt informeert over de vaardigheid van de leerling, maar eveneens beperkt informeert over beoordelingsbias van docenten of over de kwaliteit van het proefwerk. De voornaamste reden waarom docenten meerdere proefwerkcijfers dienen te verzamelen om tot betrouwbare beoordelingen te komen moet daarom gezocht worden in dit residu. Wanneer onderzoek wil nagaan hoe de becijfering efficiënter gemaakt zou kunnen worden – zodat minder cijfers nodig zijn om tot een betrouwbare beoordeling te komen – dan is het nodig om meer facetten mee te wegen dan alleen de kwaliteit van proefwerken en de beoordelingsbias van docenten.

Bovengenoemde resultaten suggereren dat de gemiddelde rapportcijfers dus redelijk vrij zijn van beoordelingsbias door de docent en docent × leerling interactie. Vervolgens is met een D-studie gepoogd een beter beeld te krijgen van de effecten op de betrouwbaarheid van beoordeling (Figuur 8). Hierin is de verwachte betrouwbaarheid voor geschiedenis weergegeven met de ononderbroken lijn, voor wiskunde met de streepjeslijn, voor Nederlands met de stippel-streepjeslijn, en voor Engels met de stippellijn. Het meewegen van de docent × leerling interactie zorgt voor een daling in de betrouwbaarheid. Deze is groter bij geschiedenis ($\Delta \boldsymbol{E}\rho^2 = .11$) en wiskunde ($\Delta \boldsymbol{E}\rho^2 = .12$), omdat de docent × leerling interactie hier een klein effect heeft (8%) dan bij Nederlands ($\Delta \boldsymbol{E}\rho^2 = .02$) en Engels ($\Delta \boldsymbol{E}\rho^2 = .02$), omdat bij de talen docent × leerling interacties verwaarloosbaar klein zijn (4 en 3%). De bovenstaande schattingen van de betrouwbaarheid van rapportcijfers geven weinig grond om te suggereren dat rapportcijfers, mits minimaal 8 proefwerken zijn gegeven, onbetrouwbaar zijn of in grote mate vertroebeld worden door bias.

Figuur 8

Grafiek van de D-studie gebaseerd op eerste geneste design (l × p): d. Op de y-as de verwachte relatieve betrouwbaarheid (***E***ρ^2). Op de x-as het aantal proefwerken bij n$_d$ = 1. De ononderbroken lijn = geschiedenis, de streepjeslijn = wiskunde, de stippel-streepjeslijn = Nederlands, en de stippellijn = Engels



5 Conclusies en Discussie

In eerder zelfrapportage-onderzoek onder docenten werd gesteld dat docenten subjectieve beoordelaars zijn die niet tot een betrouwbare beoordeling van leerlingen kunnen komen. Docenten zouden kenmerken anders dan de getoonde vaardigheid, zoals motivatie, persoonlijkheid en de doorgemaakte ontwikkeling, meewegen in het uiteindelijke cijfer waardoor leerlingen met gelijke kwaliteiten toch anders beoordeeld worden (hodgepodge-hypothese). Ook zouden docenten verschillen in mildheid van beoordelen (mildheidshypothese). In dit verkennende onderzoek is geen overtuigend bewijs gevonden voor deze beide hypotheses. De rapportcijfers van leerlingen lijken niet in hoge mate vertroebeld doordat leerlingen van gelijke kwaliteiten anders worden beoordeeld. Verder lijken de beoordeling van leerlingen ook niet in grote mate te worden vertroebeld door verschillen in mildheid tussen docenten. We concluderen dat de rapportcijfers – mits minimaal 8 proefwerken worden gegeven in een schooljaar – redelijke betrouwbaarheid hebben voor relatieve beoordelingen ($\boldsymbol{E}\rho^2 \geq .70$) en goede betrouwbaarheid hebben voor absolute beoordelingen over de cesuur van voldoende-onvoldoende ($\boldsymbol{\Phi}_\lambda \approx .90$). Deze conclusie is in lijn met conclusies van andere recente werken waarin is geconcludeerd dat subjectiviteit in becijfering niet een zeer grote rol kan spelen gezien de correlaties tussen schoolcijfers en andere variabelen voor schoolsucces (bijv., Bowers, 2010; Südkamp, Kaiser, & Möller, 2012). Dit betekent niet perse dat het percentage docenten dat rapporteert hodgepodgegedrag te vertonen incorrect is of dat docenten overdrijven. Het is nog steeds mogelijk dat 37-39% van de docenten soms hodgepodgegedrag vertonen. De resultaten suggereren eerder dat het aantal leerlingen waarbij deze 37-39% waarbij docenten ervoor kiezen om ook anderen facetten dan de vaardigheid te laten meewegen in hun beoordeling laag is en/of dat docenten niet telkens bij dezelfde leerlingen hodgepodgegedrag vertonen. Leerlingen worden vooral beoordeeld op basis van hun vaardigheid.

We sommen hieronder enkele andere belangrijke resultaten van deze studie op: (1) Voor rapportcijfers die zijn gebaseerd op minder dan 8 proefwerkcijfers is verwachte betrouwbaarheid voor de relatieve besluiten lager dan het criterium van .70. Er zijn dus minimaal 8 proefwerken nodig om tot een consistente rangorde te komen van minst tot meest vaardig in een schoolvak; (2) De subjectiviteit in beoordeling is het hoogst bij rapportcijfers van een 7.0. Voor een 7.0 is het dus het meest onzeker of dit ook echt een 7.0 is of eigenlijk een 6.0 of een 8.0; (3) Onbetrouwbaarheid in rapportcijfers kan vooral verklaard worden door verschillen in de kwaliteit van proefwerken. De verschillen in kwaliteit van proefwerken wegen zwaarder dan de subjectiviteit van beoordelaars; (4) Er zijn relevante verschillen tussen schoolvakken in de betrouwbaarheid van beoordeling waarbij de rapportcijfers voor de talen en vooral bij Nederlands, een lagere betrouwbaarheid hebben dan de beoordelingen voor geschiedenis en wiskunde. Toch geldt ook hier dat deze lagere betrouwbaarheid niet lijkt te komen door hogere subjectiviteit in beoordeling van de vakdocenten Nederlands of Engels.

5.1 Methodologische beperkingen

Bij de interpretatie van de resultaten van deze studie moet rekening worden gehouden met een aantal beperkingen. De belangrijkste beperking is de geringe steekproefgrootte van 64 en 52 docenten. Hierdoor is het aantal docenten per schoolvak gering. De resultaten voor de mildheidshypothese konden in de beide steekproeven worden geanalyseerd en deze kruisvalidatie laat zien dat de resultaten plausibel zijn. Toch konden de resultaten voor de hodgepodgehypothese niet in beide steekproeven worden geanalyseerd en de grootte van de tweede steekproef – twee scholen en 58 docenten – beperkt de generaliseerbaarheid van de resultaten.

Een tweede beperking is dat bij deze methode opeenvolgende cijfers beschouwd worden als onafhankelijke waarnemingen. De assumptie van onafhankelijke waarnemingen wordt regelmatig geschonden, maar in het specifieke geval van schoolcijfers is onduidelijk hoe ernstig deze schending is.

De resultaten van de tweede deelstudie worden ook beperkt door de uitval (27.7%). De meeste uitvallers betroffen leerlingen die verhuisden naar een andere schoolsoort door buitengewoon hoge of lage prestaties. Uit het resultaat in Figuur 7 blijkt dat leerlingen aan de uitersten van de cijferschaal met hoge betrouwbaarheid worden beoordeeld. Met het wegvallen van deze leerlingen is de heterogeniteit in de steekproef toegenomen. Het meest logische gevolg van deze uitval is daarom dat we de beoordelingsbias hebben overschat.

Als laatste punt erkennen we dat het gebruikte criterium voor betrouwbaarheid: ($\boldsymbol{E}\rho^2 \geq .70$) voor discussie vatbaar is. Psychometrici hebben geargumenteerd dat het criterium ≥ .70 adequaat is voor explorerend en fundamenteel onderzoek, maar dat besluiten waarvan veel afhangt voor de personen in kwestie een hogere betrouwbaarheid vereisen (bijv., Nunnally, 1978). We merken op dat bij het criterium ≥ .70 ook in het geval van een 'betrouwbaar' rapportcijfer nog steeds een aanzienlijk deel van het besluit wordt bepaald door facetten anders dan de vaardigheid.

5.2 Praktische en theoretische relevantie

Het is lang bekend dat een beoordeling op basis van één enkele proefwerk kwetsbaar is, omdat de leerling een minder moment kan hebben (Spearman, 1910). Om een goed beeld te krijgen van de kwaliteiten van een leerling zijn meerdere observaties nodig, maar schoolvakken verschillen in hoge mate in de hoeveelheid proefwerken die ze geven aan de leerlingen. Uit dit onderzoek blijkt dat er minimaal 8 cijfers nodig zijn om tot een redelijk betrouwbare beoordeling te komen. Dit betekent dat er bij geschiedenis in de meeste gevallen meer cijfers nodig zijn. Voor de talen is het lastiger om op basis van dit onderzoek tot een advies te komen over het aantal proefwerken. We hebben gespeculeerd dat de lagere betrouwbaarheid bij de talen het gevolg zou kunnen zijn van de diversiteit in de (deel)vaardigheden die in dit vak worden beoordeeld. De veronderstelling is dat een leerling die bijvoorbeeld literair competent is, niet ook sterk hoeft te zijn in andere onderdelen zoals spreken of schrijven. Als deze veronderstelling klopt dan zijn er voor het vak Nederlands veel meer proefwerken nodig, omdat iedere aparte (deel)vaardigheid dan een aantal keren becijferd zou moeten worden. Wanneer deze veronderstelling niet klopt, dan kunnen de talen volstaan met 10-12 cijfers in een schooljaar.

Relevant voor de onderzoekspraktijk is dat de resultaten suggereren dat – mits er minimaal 8 cijfers zijn gegeven – de rapportcijfers een waardevolle bron van de informatie zijn over de kwaliteiten van een leerling. Dit suggereert dat de rapportcijfers bruikbaar zijn voor onderzoek naar het leren van leerlingen. Tegelijk suggereren de resultaten ook dat er weinig variatie is in rapportcijfers tussen docenten. Dit suggereert dat rapportcijfers onbruikbaar zijn voor onderzoek naar het functioneren van docenten. Andere vormen van data zijn nodig voor zulk onderzoek.

5.3 Mogelijkheden voor toekomstig onderzoek

Naast replicatie van de huidige resultaten in een grotere steekproef geven de resultaten van dit onderzoek nog andere suggesties voor vervolgonderzoek. Uit de resultaten blijkt dat de variatie tussen proefwerken groter is dan de variatie tussen beoordelaars. Vervolgonderzoek zou zich kunnen toeleggen op de vraag waarom klasgemiddeldes variëren van proefwerk op proefwerk. Er kunnen minimaal drie hypotheses worden getoetst: (1) de klasgemiddelde prestaties op proefwerken fluctueren omdat de proefwerken verschillen in moeilijkheid. Deze eerste hypothese zou erop duiden dat meerdere docenten die dezelfde proefwerken geven ook op dezelfde proefwerken lagere en hogere klasgemiddelde cijfers behalen; (2) de klasgemiddelde prestaties op proefwerken fluctueren omdat de docent in sommige onderwerpen beter lesgeeft dan in andere onderwerpen. Deze tweede hypothese zou erop duiden dat wanneer een docent klassen verschillende proefwerken geeft over hetzelfde onderwerp alle klassen voor hetzelfde lesonderwerp lagere of hogere cijfers behalen; (3) de klasgemiddelde prestaties op proefwerken fluctueren omdat de docent compenseert. Deze derde hypothese duidt op een vorm van beoordelingsbias die hier niet kon worden onderzocht. De hypothese voorspelt een hoge docent × proefwerk interactie.

Ten tweede suggereren de resultaten dat de betrouwbaarheid in beoordeling lager is bij de talen – en vooral bij Nederlands – dan bij geschiedenis en wiskunde. Vervolgonderzoek zou zich ook kunnen richten op het vakspecifieke in het betrouwbaar beoordelen van de vaardigheid bij Nederlands. In dit artikel hebben we gespeculeerd dat de lagere betrouwbaarheid in beoordeling bij het vak Nederlands deels zou kunnen liggen aan de uiteenlopende deelvaardigheden die in dit vak worden gedoceerd. Toekomstig onderzoek zou kunnen nagaan in hoeverre deze veronderstellingen gerechtvaardigd zijn en hoe hier beter mee omgegaan kan worden in de beoordeling.

Literatuur

Atkinson, R. C., & Geiser, S. (2009). Reflections on a Century of College Admissions Tests. *Educational Researcher, 38,* 665-667. DOI: 10.3102/0013189X09351981

Bates, D., Maechler, M., Bolker, B., Walker, S. (2014). *lme4: Linear Mixed-Effects Models Using S4 Classes*. R package version 1.1-7. URL: http://CRAN.R-project.org/package=lme4

Biesta, G. J. J. (2012). *Goed onderwijs en de cultuur van het meten.* Den Haag: Boom Lemma.

Bowers, (2009). Reconsidering grades as data for decision making: More than just academic knowledge. *Journal of Educational Administration, 47*, 609-629. DOI: 10.1108/09578230910981107

Bowers, (2010). Grades and graduation: A longitudinal risk perspective to identify student dropouts. *Journal of Educational Research, 103,* 191-207. DOI: 10.1080/00220670903382970

Bowers, (2011). What's in a grade? The multidimensional nature of what teacher-assigned grades assess in high-school. *Educational Research and Evaluation: An International Journal on Theory and Practice, 17,* 141-159. DOI: 10.1080/13803611.2011.597112

Brennan, R. L., (2001). *Generalizability Theory: Statistics for Social Science and Public Policy.* NY: Springer-Verlag, Inc

Brennan, R. T., Kim, J., Wenz-Gros, M., & Siperstein, G. M. (2001). The relative equitability of high-stakes testing versus teacher-assigned grades: An analysis of the Massachusetts comprehensive assessment system (MCAS). *Harvard Educational Review, 71,* 173-216.

Brookhart, S. M. (1994). Teachers' Grading: Practice and Theory. *Applied Measurement in Education, 7,* 279-301.

Brookhart, S. M. (2004). Classroom assessment: Tensions and intersections in theory and practice. *Teacher College Record, 106,* 429-458.

Cliffordson, C. (2008). Differential prediction of study success across academic programs in the Swedish context: The validity of grades and tests as selection instruments for higher Education. *Educational Assessment, 13*, 56-75. DOI: 10.1080/10627190801968240

Cronbach, L. J., Gleser, C. G., Rajaratnam, N., & Nanda, H. (1972). *The dependability of behavioral measurements.* New York: Wiley.

Cross, L., H., & Frary, R. B. (1999). Hodgepodge grading: Endorsed by teachers and students alike. *Applied Measurement in Education, 12,* 53-72.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Abe, H., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. Journal of Statistical Software, 39, 1–25.

De Groot, A. D., & Wijnen, W. H. F. W. (1983). *Vijven en zessen. Cijfers en beslissingen: het selectieproces in ons onderwijs (10de druk).* Groningen, Nederland: Wolters Noordhoff

Drany, K., & Wilson M. (2008). An LLTM approach to the examination of teachers' ratings of classroom assessment tasks. *Psychology Science Quarterly, 50*, 417-432.

Hattie, J. (2009). *Visible learning. A synthesis of over 800 meta-analyses relating to achievement.* New York, NY: Routledge

Hofstee, W., K., B. (1999). *Principes van beoordeling: Methodiek en ethiek van selectie, examinering en evaluatie.* Amsterdam, Nederland: Swets & Zeitlinger

Kuhlemeier, H., & Kremers, E. (2013). *De praktijk van de eerste en tweede correctie. Samenvatting van onderzoek naar het functioneren van het CSE.* Arnhem: Cito

Korobko, O., B., Glas, C. A. W., Bosker, R. J., & Luyten, J., W. (2008). Comparing the difficulty of examination subjects with item response theory. *Journal of Educational Measurement, 45,* 139-157.

Marzano (2002). A comparison of selected methods of scoring classroom assessments. *Applied Measurement in Education, 15,* 249-268. DOI: 10.1207/S15324818AME1503_2

McMillan, J. H., Myran, S., & Workman, D. (2002). Elementary Teachers' Classroom Assessment and Grading Practices, *The Journal of Educational Research, 95*, 203-213. DOI: 10.1080/00220670209596593

Muijs, D. (2006). Measuring teacher effectiveness: Some methodological reflections. *Educational Research and Evaluation: An International Journal on Theory and Practice, 12,* 53–74.

Nunnally, J. C. (1978). *Psychometric Theory (2nd edition).* NY: McGraw-Hill.

Randall, J., & Engelhard, G. (2010). Examining the grading practices of teachers. *Teaching and Teacher Education, 26*, 1372–1380. DOI: 10.1016/j.tate.2010.03.008

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability Theory: A Primer.* Thousand Oaks, California: Sage Publications, Inc

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 271-295.

Standaert, R. (2014). *De becijferde school. Meetcultus en meetcultuur.* Leuven, België: Acco

Starch & Elliot, (1914). Reliability of grading work in mathematics. *The school review, 21,* 254-259.

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgements of students' academic achievement: a meta-analysis. *Journal of Educational Psychology, 104,* 743-762.

Thorsen, C., & Cliffordson, C. (2008). The predictive validity of teacher-assigned criterion-referenced grades. *Educational Research and Evaluation: An International Journal on Theory and Practice,18,* 153-172. DOI: 10.1080/13803611.2012.659929

Wright, S., P., Horn, S., P., & Sanders, W., L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of personnel evaluation in education, 11,* 57-67.

APPENDIX B

ICALT Classroom observation instrument

Appendix B

The 32-item ICALT-Classroom observation instrument studied in Chapter 2, 4, 5, and 6.

Please observe the following teaching practices:

To score, please encircle your answer:

1= predominantly weak; 2=more weak than strong 3= more strong than weak; 4= predominantly strong

		This Teacher…	
Safe and stimulating learning climate	1	...shows respect for learners in his/her behaviour and language	1 2 3 4
	2	...maintains a relaxed atmosphere	1 2 3 4
	3	...promotes learners' self-confidence	1 2 3 4
	4	...fosters mutual respect	1 2 3 4
Efficient organisation	5	...ensures the lesson proceeds in an orderly manner	1 2 3 4
	6	...monitors to ensure learners carry out activities in the 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
	10	...gives feedback to learners	1 2 3 4
	11	...engages all learners in the lesson	1 2 3 4
	12	… during the presentation stage, checks whether learners have understood the subject material	1 2 3 4
	13	… encourages learners to do their best	1 2 3 4
	14	...teaches in a well-structured manner	1 2 3 4
	15	...gives a clear explanation of how to use didactic aids and how to carry out assignments	1 2 3 4
Intensive and activating teaching	16	... offers activities and work forms that stimulate learners to take an active approach	1 2 3 4
	17	...stimulates the building of self-confidence in weaker leaners	1 2 3 4
	18	...stimulates learners to think about solutions	1 2 3 4
	19	...asks questions which stimulate learners to reflect	1 2 3 4
	20	...lets learners think aloud	1 2 3 4
	21	...gives interactive instructions	1 2 3 4
	22	...clearly specifies the lesson aims at the start of the lesson	1 2 3 4

Adjusting instructions	23	...evaluates whether the lesson aims have been reached	1 2 3 4
	24	...offers weaker learners extra study and instruction time	1 2 3 4
	25	...adjusts instructions to relevant inter-learner differences	1 2 3 4
	26	...adjusts the processing of subject matter to relevant inter-learner differences	1 2 3 4

Teaching learning strategies	27	...teaches learners how to simplify complex problems	1 2 3 4
	28	…stimulates the use of control activities	1 2 3 4
	29	...teaches learners to check solutions	1 2 3 4
	30	...stimulates the application of what has been learned	1 2 3 4
	31	...encourages learners to think critically	1 2 3 4
	32	...asks learners to reflect on approach strategies	1 2 3 4

APPENDIX C

“My Teacher” questionnaire (59 items)

Appendix C

The “My Teacher” questionnaire used in Chapter 3 (in bold the selected 28 items used in Chapter 4). Answer categories: 1 = weak, 2 = more weak than strong, 3 = more strong than weak, and 4 = strong. The data gathered after the summer 2013 and reported upon in Chapter 4 had other answer categories: 1 = never, 2 = seldom, 3 = regularly, and 4 = often (in the data obtained in the ministry financed induction project and the ministry financed project evaluation at weak performing schools). The independent research project used a binary response format in which the answer categories were: 1 = seldom, and 2 = often.

	My teacher…	***1***	***2***	***3***	***4***
9	… ensures that I am relax in the classroom	O	O	O	O
11	… ensures that I do my best	O	O	O	O
12	**… ensures that I know the lesson goals**	O	O	O	O
13	**… involves me in the lesson**	O	O	O	O
14	… explains clearly	O	O	O	O
15	… ensures that I behave well	O	O	O	O
16	**… ensures that I use my time effectively**	O	O	O	O
17	**… keeps track of what I know and am capable of**	O	O	O	O
18	… gives extra time for tasks that I find difficult	O	O	O	O
19	… makes me feel self-confident with difficult tasks	O	O	O	O
20	… ensures that I cooperate well with peers	O	O	O	O
21	**… treats me with respect**	O	O	O	O
22	… checks whether I reached the lesson goal	O	O	O	O
23	**… explains everything clearly to me**	O	O	O	O
24	… asks questions that make me thinking	O	O	O	O
25	**… checks whether I understood the subject matter**	O	O	O	O
26	**… answers my questions**	O	O	O	O
28	… ensures that I am active during the lesson	O	O	O	O
30	… checks whether I perform the assignments well	O	O	O	O
31	**… stimulates my thinking**	O	O	O	O
32	… teaches me to think aloud	O	O	O	O
33	**… teaches me to check my own solutions**	O	O	O	O
34	… teaches me to simplify problems	O	O	O	O
35	… explains why the material taught is useful	O	O	O	O

36	… makes me check my own work for mistakes	O	O	O	O
37	**… encourages me to think for myself**	O	O	O	O
38	**… ensures that others treat me with respect**	O	O	O	O
39	**… ensures that I pay attention**	O	O	O	O
40	**… ensures that I keep working**	O	O	O	O
41	… quickly notices when I misbehave	O	O	O	O
43	**… ensures that I know what to do**	O	O	O	O
44	**… ensures that I treat others with respect**	O	O	O	O
45	**… explains the purpose of the lesson**	O	O	O	O
46	… has high demands for me	O	O	O	O
49	**… connects to what I know or am capable of**	O	O	O	O
50	… notices when I do not pay attention	O	O	O	O
51	… does give notice to me if I improve	O	O	O	O
52	… teaches me to ask myself questions when I am reading	O	O	O	O
55	… let me work in my own pace	O	O	O	O
56	**… makes clear when I should have finished an assignment**	O	O	O	O
57	**… applies clear rules**	O	O	O	O
58	… has attention for me	O	O	O	O
59	… notices when something is disturbing me	O	O	O	O
60	… intervene when I disturb the order	O	O	O	O
65	**… knows what I find difficult**	O	O	O	O
66	**… teaches me to summarize what I have read in my own words**	O	O	O	O
68	**… prepares his/her lesson well**	O	O	O	O
69	**… makes clear what I need to study for a test**	O	O	O	O
70	… ensures that I understand the explanation	O	O	O	O
71	… offers help when I don't know or can do something	O	O	O	O
72	… encourages me to think for myself	O	O	O	O
73	**… evokes interest**	O	O	O	O
74	… motivates me	O	O	O	O
76	… offers examples what can be done with the lesson material	O	O	O	O
78	**… uses clear examples**	O	O	O	O

79	**… helps me if I do not understand or am unable to do something**	O	O	O	O
81	**… explains me how I can do something**	O	O	O	O

APPENDIX D

ICALT- Lesobservatie-instrument

Appendix D

Het 32 item ICALT lesobservatie-instrument dat is gebruikt in Hoofdstuk 2, 4, 5, en 6.

Observeer de volgende gebeurtenissen:

Oordeel: Omcirkel s.v.p. het gewenste antwoord:

1= overwegend zwak; 2=meer zwak dan sterk 3= meer sterk dan zwak; 4= overwegend sterk

		Indicator: De leraar ...	Oordeel
Veilig en stimulerend leerklimaat	1	...toont in gedrag en taalgebruik respect voor leerlingen	1 2 3 4
	2	...zorgt voor een ontspannen sfeer	1 2 3 4
	3	...ondersteunt het zelfvertrouwen van leerlingen	1 2 3 4
	4	...zorgt voor wederzijds respect	1 2 3 4
Efficiënte lesorganisatie	5	...zorgt voor een ordelijk verloop van de les	1 2 3 4
	6	...gaat tijdens de verwerking na of leerlingen de opdrachten op een juiste manier uitvoeren	1 2 3 4
	7	...zorgt voor een doelmatig klassenmanagement	1 2 3 4
	8	...gebruikt de leertijd efficiënt	1 2 3 4
Duidelijke en gestructureerde instructie	9	...geeft duidelijke uitleg van de leerstof	1 2 3 4
	10	...geeft feedback aan de leerlingen	1 2 3 4
	11	...betrekt alle leerlingen bij de les	1 2 3 4
	12	...gaat tijdens de instructie na of leerlingen de leerstof hebben begrepen	1 2 3 4
	13	...bevordert dat leerlingen hun best doen	1 2 3 4
	14	...geeft goed gestructureerd les	1 2 3 4
	15	...geeft duidelijke uitleg van het gebruik van didactische hulpmiddelen en opdrachten	1 2 3 4
Intensieve en activerende les	16	...hanteert werkvormen die leerlingen activeren	1 2 3 4
	17	...stimuleert het zelfvertrouwen van zwakke leerlingen	1 2 3 4
	18	...stimuleert leerlingen om over oplossingen na te denken	1 2 3 4
	19	...stelt vragen die leerlingen tot denken aanzetten	1 2 3 4
	20	...laat leerlingen hardop denken	1 2 3 4
	21	...zorgt voor interactieve instructie	1 2 3 4
	22	...verduidelijkt bij de aanvang van de les de lesdoelen	1 2 3 4

Afstemmen van instructie en verwerking op verschillen	23	...gaat na of de lesdoelen werden bereikt	1 2 3 4
	24	...biedt zwakke leerlingen extra leer- en instructietijd	1 2 3 4
	25	...stemt de instructie af op relevante verschillen tussen leerlingen	1 2 3 4
	26	...stemt de verwerking van de leerstof af op relevante verschillen tussen leerlingen	1 2 3 4

Leerstrategieën aanleren	27	...leert leerlingen hoe zij complexe problemen kunnen vereenvoudigen	1 2 3 4
	28	…stimuleert het gebruik van controle activiteiten	1 2 3 4
	29	...leert leerlingen oplossingen te checken	1 2 3 4
	30	...bevordert het toepassen van het geleerde	1 2 3 4
	31	...moedigt kritisch denken van leerlingen aan	1 2 3 4
	32	...vraagt leerlingen na te denken over strategieën bij de aanpak	1 2 3 4

Het is niet toegestaan om de inhoud van dit observatie instrument te veranderen, kopiëren, aan te vullen of te herschrijven zonder toestemming van Prof. Wim van de Grift, dr. Michelle Helms-Lorenz, of dr. Ridwan Maulana. Wanneer u interesse heeft in het gebruiken van de vragenlijst kan u contact opnemen met de Lerarenopleiding van de Rijksuniversiteit Groningen.

APPENDIX E

“Mijn leraar…” vragenlijst

(59 items)

Appendix E

De “Mijn Leraar…” vragenlijst gebruikt voor Hoofdstuk 3. Dikgedrukt staan de 28 items die zijn meegenomen in het onderzoek in Hoofdstuk 4. De leerlingen hadden oorspronkelijk 4 antwoordopties: 1 = zwak, 2 = meer zwak dan sterk, 3 = meer sterk dan zwak, en 4 = sterk. De data die is verzameld na de zomer van 2013 en waarover wordt gerapporteerd in Hoofdstuk 4 had andere antwoordopties 1 = zelden, 2 = vaak.

	Mijn leraar…	***1***	***2***	***3***	***4***
9	… zorgt dat ik in de klas ontspannen ben	O	O	O	O
11	… zorgt dat ik mijn best doe	O	O	O	O
12	**… zorgt dat ik weet wat lesdoelen zijn**	O	O	O	O
13	**… betrekt mij bij de les**	O	O	O	O
14	… geeft mij goed les	O	O	O	O
15	… zorgt dat ik mij goed gedraag	O	O	O	O
16	**… zorgt dat ik mijn tijd goed gebruik**	O	O	O	O
17	**… houdt in de les rekening met wat ik al weet of kan**	O	O	O	O
18	… geeft mij extra tijd voor taken die ik moeilijk vind	O	O	O	O
19	… geeft mij zelfvertrouwen bij alle moeilijke taken	O	O	O	O
20	… zorgt dat ik goed samenwerk met anderen	O	O	O	O
21	**… benadert mij met respect**	O	O	O	O
22	… controleert of ik de lesdoelen heb bereikt	O	O	O	O
23	**… legt mij alles duidelijk uit**	O	O	O	O
24	… stelt mij vragen waarover ik moet nadenken	O	O	O	O
25	**… controleert of ik de lesstof heb begrepen**	O	O	O	O
26	**… geeft antwoord op mijn vragen**	O	O	O	O
28	… zorgt dat ik in de les actief ben	O	O	O	O
30	… controleert of ik de opdrachten goed uitvoer	O	O	O	O
31	**… stimuleert mij om na te denken**	O	O	O	O
32	… laat mij hardop denken	O	O	O	O
33	**… leert mij mijn oplossingen te controleren**	O	O	O	O
34	… leert mij problemen te vereenvoudigen	O	O	O	O
35	… vertelt waarvoor ik de lesstof kan gebruiken	O	O	O	O

36	… laat mij mijn eigen werk controleren	O	O	O	O
37	**… moedigt mij aan om zelf na te denken**	O	O	O	O
38	**… zorgt dat anderen mij met respect behandelen**	O	O	O	O
39	**… zorgt dat ik oplet**	O	O	O	O
40	**… zorgt dat ik doorwerk**	O	O	O	O
41	… heeft het snel in de gaten als ik mij misdraag	O	O	O	O
43	**… zorgt dat ik weet wat ik moet doen**	O	O	O	O
44	**… let erop dat ik anderen met respect behandel**	O	O	O	O
45	**… maakt mij duidelijk wat de bedoeling is van de les**	O	O	O	O
46	… stelt hoge eisen aan mij	O	O	O	O
49	**… sluit aan bij wat ik weet of kan**	O	O	O	O
50	… heeft het in de gaten als ik niet oplet	O	O	O	O
51	… laat het me merken als ik voorruit ga	O	O	O	O
52	… leert me tijdens het lezen vragen aan mezelf te stellen	O	O	O	O
55	… zet mij aan het denken	O	O	O	O
56	**… vertelt mij hoe ik iets moet leren**	O	O	O	O
57	**… laat mij in mijn eigen tempo werken**	O	O	O	O
58	… zegt duidelijk wanneer de opdracht klaar moet zijn	O	O	O	O
59	… hanteert duidelijke regels	O	O	O	O
60	… heeft aandacht voor mij	O	O	O	O
65	**… ziet het als me iets dwars zit**	O	O	O	O
66	**… grijpt in als ik orde verstoor**	O	O	O	O
68	**… weet wat ik moeilijk vind**	O	O	O	O
69	**… leert me wat ik gelezen heb in mijn eigen woorden samen te vatten**	O	O	O	O
70	… bereidt zijn/haar lessen goed voor	O	O	O	O
71	… maakt duidelijk wat ik voor een proefwerk moet leren	O	O	O	O
72	… zorgt dat ik de uitleg snap	O	O	O	O
73	**… helpt mij als ik iets niet weet of kan**	O	O	O	O
74	… moedigt mij aan om zelf na te denken	O	O	O	O
76	… vertelt boeiend	O	O	O	O

78	**… motiveert mij**	O	O	O	O
79	**… geeft voorbeelden van wat je met de lesstof kunt doen**	O	O	O	O
81	**… gebruikt duidelijke voorbeelden**	O	O	O	O

Appendix F

Technical appendix

In this appendix, we elaborate on the mathematics behind the Generalizability in Item Response Theory (GIRT) analysis method (used in Chapter 5). We note that further explanation of the concepts, statistical theory and the mathematics is provided by Briggs and Wilson (2007) and Choi (2012) and we only briefly summarize this here.

From the conceptual point of view, GIRT attempts to combine the need for measurement invariance accentuated by IRT, with the need for controlling for dependencies (or biases) due to non-random sampling accentuated by Generalizability theory[4]. This combination has much potential, because it provides a sample invariant metric which is tested for certain properties – in our case cumulative item order (see section 1.5) – and then examines how group membership (e.g., teaching a specific class, being observed by specific persons, or teaching a specific subject) affects teachers' position on this metric. Furthermore, it has the potential to investigate whether the bias due to group membership makes that the by IRT produced personal evaluation scores are too unreliable. If this is the case, then we cannot use the personal evaluation scores, despite that at the sample level we see low standard errors of measurement and/or good model fit.

Mathematics behind GIRT

To combine Generalizability theory with the Rasch model, we first need to transform the original Rasch model into a generalized linear mixed model (GLMM). After this is done, facets can be added to this equation to retrieve information about group membership. There are multiple ways how to transform Item Response Theory models into a GLMM (e.g., de Boeck et al., 2011; Choi, 2012; Fox, 2010). However, in Chapter 5 we followed the approach by Choi (2012) and in this dissertation we use the notation as applied by de Boeck et al. (2011).

The Rasch model estimates the chance on a correct response using a log-odds scale. To transform the chance on a correct response (P) into the log-odds, it is divided by the change on failure ($1 - P$) and then we take to logarithm of this. This part left of the "equals" sign in the equation (1) is identical to an ordinal regression.

[4] in non-psychometrical jargon the statistical analyses resulting from Generalizability theory are often referred to as (cross-classified)-multilevel models. The reader may read multilevel analysis instead of Generalizability theory, if (s)he is more acquainted with this technique.

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

The subscript “ti” signifies that the chance of a correct response is dependent on two facets, namely t and i. The t refers to the parameter θ_t, which describes the differences in teaching we observe, and the i refers to the parameter b_i which describes the differences in complexity of the teaching practices we are observing. The parameter θ_t is connected to the dummy variable *Z* (as in a regression model *b**dummy). *Z* is a covariate coded 1 if *j* = t and 0 otherwise, in which *j* is the sum score of the teacher. Thus, the *Z*-dummy groups teachers of equal skill together. This obliges the model to seek for estimates of item complexity which show the least variation between these groups. The parameter b_i is connected to the dummy covariate *X*, which adapts each item score to the point where *k* = i, in which *k* is the sum score of an item. Here, the same logic applies. The dummy variables are accompanied by the subscript (t, i) to reflect that both parameters need to be extracted from the same single data point. This part (t, i) may be read as a direction to search for the particular cells in the data matrix that match a unique teacher (t) with an unique item (i). In this specific case, this is only one cell. As de Boeck, et al. (2011) explain, the parameters $\theta_t Z_{(t,i)j}$ and $b_i X_{(t,i)k}$ may be specified as ‘fixed’ or ‘random’ effects depending on how one would like to estimate the model. For the research presented in Chapter 5, the parameters were specified to be random effects. In subsequent notation, we therefore also use, for example, “σ_t^2” to refer to the random parameter $\theta_t Z_{(t,i)j}$.

De Boeck et al. (2011) describe how to decompose the facets further. The G-study (design o × t × I mentioned in Figure 5.5 Chapter 5 would estimate the variances in each of the six observed facets: t, o, i, to, ti, and oi where the “:” should be read as “nested in” and letter combinations are interactions. The multilevel Rasch model equation then is:

$$\log\left(\frac{P_{\mathrm{oti}}}{1\text{-}P_{\mathrm{oti}}}\right) =$$

$$\sum_{\mathrm{j}=1}^{\mathrm{J}} \theta_{\mathrm{t}} Z_{(\mathrm{t,i})\mathrm{j}} + \sum_{\mathrm{l}=1}^{\mathrm{L}} \theta_{\mathrm{o}} V_{(\mathrm{o,t})\mathrm{l}} \sum_{\mathrm{j}=1}^{\mathrm{J}} \sum_{\mathrm{l}=1}^{\mathrm{L}} \theta_{\mathrm{ot}}\ W_{((\mathrm{to}),\mathrm{i})\mathrm{jl}} + \sum_{\mathrm{k}=1}^{\mathrm{K}} b_{\mathrm{i}} X_{(\mathrm{ot,i})\mathrm{k}} +$$

$$\sum_{\mathrm{j}=1}^{\mathrm{J}} \sum_{\mathrm{k}=1}^{\mathrm{K}} \theta_{\mathrm{t}} Z_{(\mathrm{t,i})\mathrm{j}}\, b_{\mathrm{i}} X_{((\mathrm{ot}),\mathrm{i})\mathrm{k}} + \sum_{\mathrm{j}=1}^{\mathrm{J}} \sum_{\mathrm{k}=1}^{\mathrm{K}} \theta_{\mathrm{o}}\ V_{(\mathrm{o,i})\mathrm{l}}\ b_{\mathrm{i}} X_{((\mathrm{ot}),\mathrm{i})\mathrm{k}}\ , (2)$$

where J defines the teachers $j_1 \dots j_n$. The covariate Z adapts the personal scores to the point where j = t. Furthermore, L defines the observers $l_1 \dots l_n$, and V is a covariate that adapts the scores for each observer to the point where l = o. Finally, K defines the items $k_1 \dots k_n$. Because items are crossed with observers (o) and teachers (t), the two interaction facets defined by the matrices $\mathbf{S}(\theta Z \times bX)$ and $\mathbf{S}(\theta V \times bX)$ also should be modeled. However, these interactions indicate violations or imperfections in the measurement model.

The lme4 syntax used to estimate the facets in Equation 2 is as follows:

```
library(lme4)
> # GIRT
> GEN.DATA=read.table("I://RUG//Onderzoek Promotie//Data//Data
> 2015//Lesobservaties//150812 hierarchisch selectie leraren MS1-2.csv",
> header=T, sep= ",")
> View(GEN.DATA)
> Data.lesob$Itemnumber.f = factor(Data.lesob$Itemnumber) # correct identification of items codes
> Data.lesob$Teacher.f = factor(Data.lesob$Teacher) # correct identification of teachers’ codes
> Data.lesob$Observer.f = factor(Data.lesob$observator) # correct identification observer codes
> Data.lesob$Case.f = factor(Data.lesob$ï..id) # correct identification of teacher × observer
interaction

# Formula 2
> Gstudy = glmer(Score ~ 1 + (1 | Teacher.f) + (1 | Observer.f) + (1 | Observer.f:Teacher.f) +
>             (1 | Itemnumber.f) + (1 | Itemnumber : Teacher.f) + (1 | Itemnumber : Observer.f),
>             data=Data.lesob, binomial)
> summary(Gstudy) # results reported in Table 1
```

The D-study design was (l:(o × t)) × I. Here, the items are treated as fixed, because the Rasch (1960) model assumptions hold, so the items have an approximately fixed and invariant rank ordering that is identical for each teacher. Therefore, the variance due to items is omitted from the formula 3. Reliability was estimated as:

$$\boldsymbol{E}\rho^2 = \frac{\sigma^2_{\mathrm{t}}}{(\sigma^2_{\mathrm{t}} + (\sigma^2_{\mathrm{o}} / n_{\mathrm{o}}) + (\sigma^{2}_{\mathrm{ot}} / n_o) + (\sigma^2_{(\mathrm{tI})} / n_{\mathrm{I}}) + (\sigma^2_{(\mathrm{oI})} / n_{\mathrm{I}}\, n_{\mathrm{o}}) + ((2\pi / 3) / n_o n_{\mathrm{I}}))}, (3)$$

where σ_t^2 defines the variance in teachers' teaching skill, σ_o^2 defines the variance among observers, σ_{ot}^2 defines the variance due to interactions between observers and teacher, $\sigma_{(tI)}^2$ defines the variance due to interactions between items and teachers, and $\sigma_{(oI)}^2$ defines the variance due to interactions between items and observers. These subscripts correspond to subscripts in Equation 2. For more information about the final element $(2\pi/3)/n_o n_I$, we refer readers to Choi (2012). To estimate the effect of additional observers (as reported in Figure 5.6), we varied the number n_o keeping everything else fixed.

References

Briggs, D. C., & Wislon, M. (2007). Generalizability in Item response theory. *Journal of Educational Measurement, 44*, 131 – 155.

Choi, J. (2013). *Advances in combining generalizability theory and item response theory.* Doctoral dissertation, University of California, Berkeley.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Abe, H., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software, 39,* 1–25.

Fox, J-P. (2010). *Bayesian Item Response Modeling. Theory and Applications.* New York: USA, Springer

Dankwoord

Tijdens het schrijven heb ik me best wel eens eenzaam gevoeld. Toch kan ik niet zeggen dat ik alleen aan het werk was. Ik wil hierbij zonder namen te noemen alle docenten, schoolleiders, roostermakers en ander administratief personeel dat betrokken is geweest bij het verzamelen van ruim 200 lesobservaties, een kleine 100 leerlingenvragenlijsten, en duizenden schoolcijfers bedanken. Ik weet dat het voor jullie een ontzettend karwei was om 3 lesbezoeken te plannen op tijdstippen dat de betreffende collega lesgaf aan de juiste klas of om alle cijfers van een klas van drie opeenvolgende jaren uit de administratie te krijgen. Ook wil ik in het algemeen de collega's van de lerarenopleiding van de Rijksuniversiteit en die daarbuiten bedanken. Sommige van jullie hebben naar eerdere versies van teksten gekeken en hier feedback op gegeven waarvoor ik erg dankbaar ben. Anderen hebben lesobservaties van mij overgenomen, mij geintroduceerd op scholen of data beschikbaar gesteld voor analyse. En met velen van jullie heb ik wel eens een kop koffie, thee of een biertje gedronken, de stress eraf gelachen en inspiratie opgedaan. Van mijn directe collega's op de lerarenopleiding wil ik naast mijn promotoren er twee in het bijzonder noemen, omdat ik tijdens het project altijd op hun steun heb kunnen rekenen.

Tim Huijgen was gedurende deze 4 jaar mijn kantoorgenoot en hij zat dus altijd in de vuurlinie als er weer eens iets niet helemaal ging zoals gepland (en dat was vaak). Vanaf het begin af aan hebben we goed met elkaar kunnen opschieten en ik heb bijzonder veel steun gehad aan je. Er was altijd ruimte om ideeën te testen en ook toen ik je na de zoveelste keer vroeg of je "even tijd had om deze alinea te lezen" deed je dat nog. Hiervoor mijn grote dank!

Marjon Fokkens-Bruinsma heeft alle ups en downs in mijn promotietraject meegemaakt. Jouw steun en feedback heeft me gemotiveerd door te zetten op de momenten dat het moeizaam ging. Als positivist heb je me altijd voorgehouden dat ik niet "zo pessimistisch hoefte te zijn". Er komt altijd weer een andere keer en uiteindelijk lukt het wel. En wat heb je gelijk gekregen! De tekst van "once is not enough" (hoofdstuk 5) heb je waarschijnlijk net zo vaak of misschien zelfs vaker gelezen en becommatarieert dan Wim en Klaas. Ik heb je professionele advies altijd zeer gewaardeerd en hoop er in de toekomst ook nog soms gebruik van te mogen maken. Veel dank aan jou dus!

Wim van de Grift is professor die me aannam als promovendus en wat ben je belangrijk en bepalend geweest in mijn professionele ontwikkeling. Toen ik binnenkwam bij jou had ik weinig tot geen kennis van psychometrie en had nog nooit met IRT of het Rasch model gewerkt. Dat ligt nu wel anders. Met jouw energie en wilskracht dwong je me altijd weer om nog een stapje verder te zetten. Mijn onderzoeksplannen waren niet gemakkelijk uit te voeren: het plan voor de dataverzameling was ambiteus en de beoogde statistische analyses vereisten een combinatie van psychometrische technieken die nog bijna nooit waren gecombineerd. Jij noemde me dan ook met enige regelmaat "een gek, maar wel eentje in de goede zin van het woord!", zo lachte je dan. Vanaf dag één heb je me altijd 100% gesteund in al onze "intellectuele avonturen" zoals je ze zelf noemde. Vrijwel nooit leek je te twijfelen of we samen de complexiteit wel aan zouden kunnen. Ik weet dat ook jij van deze gesprekken genoten hebt. Dank voor alles.

Klaas van Veen is de initiator geweest, de "vader" zoals Wim het wel noemde. Met jouw rust, relativeringvermogen en toch ook rotsvaste geloof ben je altijd ontzettend belangrijk geweest voor mij. Jij was degene die mij uiteindelijk aannam na mijn masterstudie. Ik had het geloof op een baan aan de universiteit eigenlijk al opgegeven toen je me toch de kans gaf aan het ICLON te komen werken. Ik heb leren schrijven en uitleggen van jou, vooral tijdens het tweede jaar van mijn promotie. Jij vond dat "ongeacht de complexiteit van het onderwerp, het personeel en de gasten van de lawaaige kroeg op de hoek moest kunnen begrijpen waar mijn onderzoek over ging en waarom het belangrijk was". Daarmee legde je de lat niet bepaald laag. Ik weet nog dat ik aan het eind van mijn tweede jaar met een journalist van Didactief praatte op de borrel van de ORD. Ze vroeg me waar mijn onderzoek over ging. Na mijn uitleg zij ze vrolijk: "je bent zeker net begonnen, als je wat verder komt wordt het vanzelf duidelijker wat je wilt". Weinig keren heb ik zo erg het gevoel gehad dat ik had gefaald. Gelukkig was je er om me weer op te lappen. Dank voor alles.

Ook mijn vrienden en familie wil ik bedanken. Mijn ouders, zus, broers en zwager hebben mij vaak gesteund. Mijn broer Tim en zijn verloofde Celine waar altijd een kamer vrij was in hun appartement in Groningen. Het logeren bij jullie vond ik altijd gezellig ook al kwam iedereen pas rond 20.00 uur thuis en lagen we om 21.30 uur toch vaak alweer op bed. Mijn zwager Wilco, voor het maken van enkele visualisaties waaronder ook de omslag van dit boek. Luuk Slooter is belangrijk geweest voor mij als rolmodel (jouw promotie lag altijd voor die van mij). Ik hoop dat je de Veni toegewezen krijgt, zodat ik je weer verder kan "volgen". Mijn ouders moet ik ook zeker bedanken. Jullie zijn er altijd voor mij en hebben altijd een luisterend oor. Als laatste wil ik Mariëtte van der Lans-Hosemans noemen. Mijn lieve vrouw; je hebt een baan in Amsterdam gecombineerd met een huishouden en onze zoon Mart, terwijl je man 3 dagen per week in Groningen verbleef. In die enigszins hectische woon-werk situatie was jij altijd degene die overzicht bewaarde en alles georganiseerd hield. Nog steeds ben je degene die me helpt om de juiste prioriteiten te zetten. De eerlijkheid is dat er zonder jou nooit iets van terrecht was gekomen.

Open brief aan de docent

"Je moet goed weten voor wie je het doet, want er komt een tijd dat het zwaar is." Ik zat aan het einde van mijn tweede jaar van mijn promotieonderzoek en luisterde naar een lezing over "leren promoveren" op de eerste pre-conference van de Onderwijs Research Dagen (ORD). Ik kan het zeker beamen, er was een tijd dat het zwaar was. Dus waarvoor doe je dan het dan? Omdat je iets aan jezelf wilt bewijzen? Omdat je dat papiertje nodig hebt? Omdat je geen alternatief weet? Allemaal redenen die je in een promotietraject kunnen houden, en ze gelden ten dele ook voor mij, maar het zijn niet redenen waar je harder voor gaat werken. Een belangrijke reden waarom ik onderzoek bleef doen waren de docenten en hun motivatie om goed les te geven. Omdat ik denk dat het thema 'lerarenevaluatie' iets belangrijks kan bijdragen aan hun werk en professie.

Evaluatie heeft voor sommigen een negatieve bijklank. Evaluatie toont ook je mindere kant en maakt daarmee kwetsbaar. Voor leraren is hun stijl van lesgeven hen dierbaar en evaluatie van lesgeven kan voor hen dan ook emotioneel zijn, zeker als het niet zo positief is als verwacht. Aan de andere kant kan evaluatie ook positiever uitvallen waardoor je tijdelijk 'onkwetsbaar' bent. Dat is de zonnige kant van evaluatie. Het kan bevestiging geven dat je 'het goed doet' en geeft jou daarmee een weerwoord tegen anderen die graag zouden zien dat jij je stijl van lesgeven wijzigt.

In de schoolpraktijk speelt evaluatie een centrale rol in de verdeling van tijd en geld, beide schaars aanwezige middelen in een school. Docenten die positief geëvalueerd worden, krijgen professionele vrijheid – ze mogen in grotere mate naar eigen inzicht hun lessen inrichten, meedoen aan onderzoek of zelf onderzoek uitvoeren –, docenten die negatief geëvalueerd worden, wordt deze professionele vrijheid tijdelijk ontzegd. Ik illustreer hoe dit kan werken in scholen met onderstaande anekdote.

> *"In mijn eerste jaar op mijn nieuwe school kwam de rectrix twee keer kijken in de les. Ze concludeerde daarna: "Kees, jij kan goed lesgeven, de klas is rustig en de les georganiseerd. Zolang de rendementen goed zijn kun je lesgeven zoals je wilt.""*

De rectrix houdt nog een voorbehoud "zolang de rendementen goed zijn," maar geeft op basis van haar evaluatie de docent professionele vrijheid en het vertrouwen om de lessen naar eigen inzicht in te richten. Overigens is mijn persoonlijk ervaring dat dit voorbehoud niet zo zwaar weegt en dat de evaluatie vooral afhangt van de lesbezoeken. Dit blijkt ook uit het vervolg van de anekdote van Kees.

> *"Na een paar jaar trad een nieuwe directie aan op de school. Ze hadden een andere pedagogisch-didactische visie op onderwijs. Na de lesbezoeken kreeg ik te horen dat*

mijn stijl van lesgeven niet paste in de visie van de school. Ik moest mijn lesgeven aanpassen ondanks dat mijn rendementen goed waren"

De anekdote laat zien hoe Kees van tijdelijke onkwetsbaarheid ineens kwetsbaar werd. Wat mij raakt in het verhaal is niet zozeer dat Kees nu tijdelijk zijn professionele vrijheid werd ontzegd – wanneer middelen schaars zijn moet een directie nu eenmaal nadenken aan wie ze deze middelen toekennen. Wat mij raakt in de anekdote is de basis waarop dit gebeurt. Wat mij raakt is het hoge subjectieve gehalte van de evaluaties van beide directeuren. Wat mij ook raakt is dat er geen mechanismen zijn in de school die dwingen tot objectievere evaluatie. Het is de subjectieve mening van Kees – ik doe het goed – tegenover de subjectieve mening van de rectrix – je doet het niet volgens onze visie –, waarbij de laatste altijd aan het langste eind trekt. Door het gebrek aan objectievere procedures is het mogelijk dat een directeur of teamleider lerarenevaluatie gebruikt als een machtsmiddel om een leraar zijn of haar 'wil op te leggen,' zoals Kees overkwam.

Kees staat niet alleen. Voor mijn onderzoek implementeerde ik evaluatie van leraren in meerdere scholen. Met de meeste scholen heb ik zeer constructief en plezierig samengewerkt, maar met enkele heb ik de samenwerking moeten opzeggen. Op de scholen waarmee ik de samenwerking stopte, ben ik geschrokken van de houding van sommige teamleiders en directeuren. In één enkel geval werd mij zelfs op de man af gevraagd "wat mijn mening was over het lesgeven van docent X." Want vervolgde de teamleider: "dat kon toch niet! Ik heb zelf ook zijn lessen bezocht en zo'n man zou eigenlijk geen les mogen geven, dat ben je met me eens toch?" Het gesprek was suggestief, bevooroordeeld en duidelijk bedoeld om een dossier op te bouwen over de betreffende docent. Een andere ervaring is een school waarmee ik goed samenwerkte totdat er een nieuwe interim-directeur werd aangesteld die als doelstelling had om 'maatregelen te treffen' voor enkele 'niet goed functionerende' docenten. Per direct kon ik mijn onderzoek daar stopzetten.

Bovenstaande ervaringen illustreren de risico's van implementatie van lerarenevaluatie in de praktijk. Het mag dan misschien ook niet verbazen dat leraren niet allemaal warmlopen voor lerarenevaluatie in hun school. Dat komt niet per se door de mogelijke teleurstelling van een minder dan verwacht resultaat. Leraren vinden het over het algemeen helemaal niet erg om tijdelijk kwetsbaar te zijn; zij weten als geen ander dat kwetsbaarheid nodig is om te kunnen leren en ontwikkelen. Wat ze wel erg vinden is de schijnbare willekeur. Evaluatie betekent voor hen vaak dat één directeur / teamleider / inspecteur / onderzoeker met haar of zijn subjectieve kijk op lesgeven op basis van 'één of enkele momenten' jouw kwetsbaarheid bepaald alsmede wat jij zou moeten leren. Voor leraren voelt dit onrechtvaardig en in sommige gevallen voelen leraren zich hierdoor zelfs onveilig.

Mijn vraag aan de beroepsgroep is daarom: wat kunnen leraren hieraan doen? Directeuren moeten nu eenmaal besluiten nemen over de verdeling van tijd en geld in de school en zij moeten daarvoor evalueren op basis van criteria. Zij zijn ook eindverantwoordelijke voor de kwaliteit van lesgeven in een school. Dit is een onontkoombare realiteit. Maar zijn directeuren en teamleiders bijvoorbeeld wel diegene die de lessen zouden moeten observeren? Zou het niet objectiever zijn als hun besluiten zouden zijn gebaseerd op lesobservaties van derden die zelf lesgeven en die de leerlingen kennen alsmede de andere omstandigheden in de school? In mijn ogen zou het beter zijn wanneer de beroepsgroep *zelf* haar lesgeven gaat evalueren en daarmee dus ook *zelf* gaat bepalen wie tijdelijk onkwetsbaar mag zijn wie tijdelijk kwetsbaar is en ook zelf een systematiek ontwikkeld waarin leraren elkaar helpen in tijden van kwetsbaarheid.

Ik hoop dat dit onderzoek meer is dan een intellectuele exercitie of zoals mijn promotoren het noemen mijn "proeve van wetenschappelijke bekwaamheid," maar dat het tegelijk ook een eerste grove schets geeft hoe lerarenevaluatie door de eigen beroepsgroep eruit zou kunnen zien. Een groot deel van het proefschrift betreft vraagstukken van betrouwbaarheid en validiteit. De meeste hoofdstukken behandelen abstracte materie zoals de volgordes van items op een vragenlijst en de vraag of en hoe collega-docenten met relatief weinig training in het observeren van lessen toch tot een betrouwbare evaluatie kunnen komen. In mijn ogen is beantwoording van zulke technische vragen belangrijk om een basis te leggen voor een door de beroepsgroep zelf georganiseerde lerarenevaluatie.

Tot slot, misschien zullen veel vragen na het lezen nog onbeantwoord blijven. Ik vraag hiervoor begrip. De snelheid waarmee wetenschap vooruit gaat verhoudt zich ten opzichte van de snelheid in het onderwijsveld als de schildpad en de haas. Wetenschap gaat langzaam, langzamer nog dan veel leraren en beleidsmakers denken. Om een beeld te schetsen. In 1966 kwam Coleman met zijn befaamde rapport "Equality of educational opportunity" met daarin als kernboodschap dat scholen vooral verschillen in de compositie van hun leerlingenpopulatie. Coleman stelde vast dat sommige scholen meer intelligente leerlingen hebben dan andere scholen en dat op sommige scholen meer kinderen zitten van rijke ouders dan op andere scholen en dat dit grotendeels zou verklaren waarom leerlingen op sommige scholen beter presteren dan op andere. De verschillen in het lesgeven van docenten zouden er veel minder toe doen, aldus Coleman (zie pagina 22). Mede als gevolg van deze conclusie uit 1966 is het belang van een goede docent gedurende een flink aantal jaren onvoldoende onderkend. Pas halverwege de jaren ’80 (bijna 20 jaar later) kwam de wetenschappelijk bewijsvoering van het tegendeel en pas na de millenniumwisseling (bijna 40 jaar later) is deze boodschap – mede dankzij populair wetenschappelijke werken zoals Hattie’s “Visible learning” uit 2009 – ook door het grote publiek en de politiek omarmd. Tegenwoordig is zoiets als “de lerarenagenda” bijna niet meer weg te denken uit het politieke beleid. Echter, de 40 jaar tijd tussen Coleman en “de lerarenagenda” spreekt boekdelen. Ik ben nu 4 jaar onderweg en deze dissertatie bespreekt slechts mijn eerste voorzichtige stappen.

Rikkert van der Lans