Teacher evaluation through observation

van der Lans, Rikkert

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

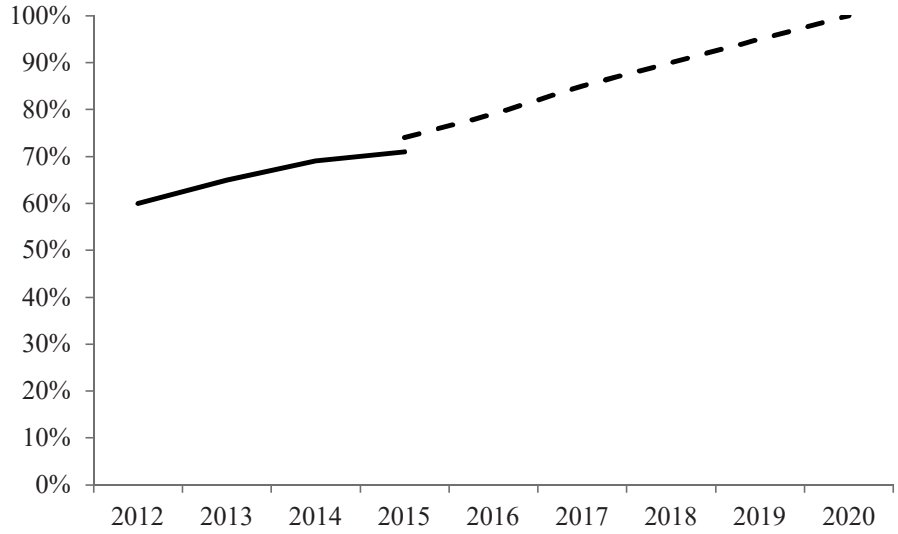
The studies addressed herein took place in the Netherlands. Organizing teacher evaluation in the Dutch context involves a relatively unique challenge compared with other countries, in that educational policies allow schools considerable autonomy in organizing teacher evaluation and grant teachers similar 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.

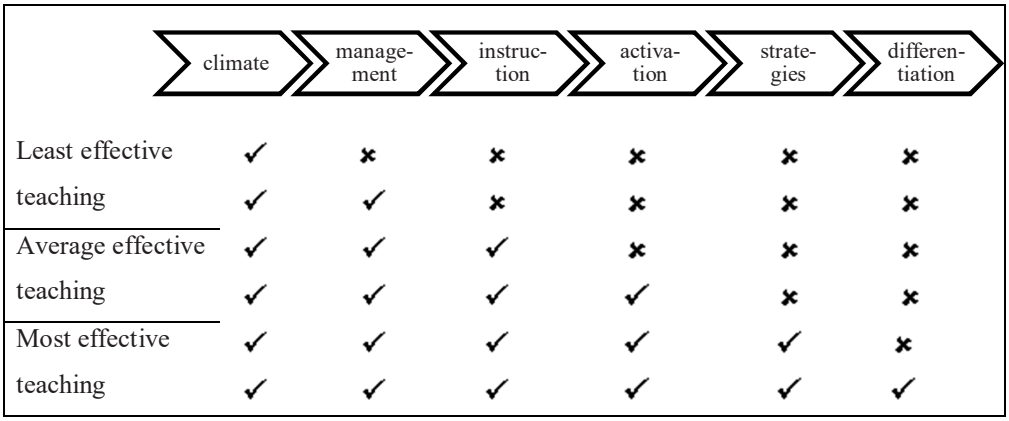
In the national induction project and the regional project focusing on low-performing schools, the instruments are interpreted as measuring teachers' development in effective teaching practices (e.g., Helms-Lorenz, Van de Grift, & Maulana, 2016; Van de Grift, Helms-Lorenz, & Maulana, 2014; Van de Grift, 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



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.