The quality and qualities of classroom observation systems

Marjoleine J. Dobbelaer
THE QUALITY AND QUALITIES OF CLASSROOM OBSERVATION SYSTEMS

Marjoleine J. Dobbelaer
THE QUALITY AND QUALITIES OF CLASSROOM OBSERVATION SYSTEMS

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof.dr. T.T.M. Palstra,
on account of the decision of the graduation committee,
to be publicly defended
on Friday, February 22, 2019 at 12.45 hours

by

Marjoleine Jolie Dobbeelaer

born on July 2, 1986
in Vlissingen, The Netherlands
This dissertation has been approved by:

Supervisor: Prof. dr. A.J. Visscher, University of Twente

Cover design: Wen Versteeg / Marjoleine Dobbelaer
Lay-out: Marjoleine Dobbelaer
Printed by: Ipskamp printing - Enschede, The Netherlands
DOI: 10.3990/1.9789036547161

© Marjoleine Dobbelaer, 2019. All rights reserved. No parts of this thesis may be reproduced, stored in a retrieval system or transmitted in any form or by any means without permission of the author.
GRADUATION COMMITTEE:

Chairman/secretary: Prof. dr. T.A.J. Toonen, University of Twente
Supervisor: Prof. dr. A.J. Visscher, University of Twente
Members:
   Prof. dr. C.A.W. Glas, University of Twente
   Prof. dr. ir. T.J.H.M. Eggen, University of Twente
   Prof. dr. P.J. den Brok, Wageningen University & Research
   Prof. dr. A.-K. Praetorius, Universität Zürich
   Dr. M. Helms-Lorenz, University of Groningen
# Table of Contents

Chapter 1  
Introduction 9

Chapter 2  
A framework in support of the development, selection, and use of classroom observation systems 19

Chapter 3  
The quality of classroom observation systems for measuring teaching quality in primary education: a systematic review 37

Chapter 4  
Quality lies in the eye of the beholder: a comparison of the perspectives on teaching quality of external raters, students, and teachers 79

Chapter 5  
Conclusion and Discussion 111

References 123

Appendix A: Evaluation framework 146

Appendix B: Search syntax in ERIC 150

Appendix C: Topics in overview document for each COS 152

Appendix D: Description and references of included COS 156

Appendix E: Impact! Items 166

Appendix F: Bugs code 168

Appendix G: Item parameter estimates and information values for three groups of raters 172

Summary in Dutch / Nederlandse Samenvatting 175

Acknowledgements / Dankwoord 187

ICO dissertation series 195
1 Introduction
1.1 Teaching quality

Student achievement is influenced by many different factors. Besides student ability, which accounts for about 50% of the variance in student achievement, home, school, principal, peer, and teacher factors also prove to have an effect (Hattie, 2003). Teachers are considered to be the greatest malleable, within-school influence on student learning (Haertel, 2013; Nye, Konstantopoulos, & Hedges, 2004). Research shows considerable variation in teachers’ impact on student achievement within schools (Nye et al., 2004). Teacher differences account for approximately 10% of the variance in student test score gains in a single year (Haertel, 2013). Results that a very effective teacher can achieve with students in half a year, can take a very ineffective teacher two years to accomplish (Visscher, 2017). Measuring teaching quality validly is therefore important, because such measures can help assess whether teacher candidates are fit for teaching (Hill, Umland, Litke, & Kapitula, 2012), guide the professional development and improvement of in-service teachers, and can also support timely and efficient human resource decisions (Haertel, 2013).

1.2 Teaching quality measures

There are several ways to measure teaching quality and there is considerable debate on what is the best method for doing so (Chetty, Friedman, & Rockoff, 2011). Goe, Bell, and Little (2008) have provided an overview of the most common methods and show that each has its own strengths and limitations. The most widely used measure is classroom observation, which can provide rich information about teacher classroom behaviors and activities, facilitating both formative (aimed at improvement) and summative (aimed at evaluation) assessment. Classroom observation is generally perceived as the most objective, fair, and direct measure (Goe et al., 2008; Lasagabaster & Sierra, 2011), however research shows that in order to obtain reliable scores, teachers need to be observed multiple times by multiple trained raters (e.g., Hill, Charalambous, & Kraft, 2012). This makes classroom observation time-consuming and expensive (Goe et al., 2008).

A more cost-effective measure can be obtained by means of a student survey for gathering student perceptions of teaching practice. Students have the most experience with the teacher and their perceptions can function as a valuable source of feedback to teachers (Bijlsma, Visscher, Dobbelaer, & Veldkamp, n.d.). However, students cannot provide information on all aspects of teaching, such as teachers’ pedagogical content knowledge. Students’ perceptions are also believed to be influenced by student
variables (e.g., ethnicity; Levy, Brok, Wubbels, & Brekelmans, 2003) and teacher features unrelated to teaching efficacy (e.g., teacher popularity; Fauth, Decristan, Rieser, Klieme, & Büttner, 2014).

A teacher survey can provide insight into teachers’ own perceptions of their teaching quality and their intentions, thought processes, knowledge, and beliefs. Teachers also have, in contrast to external raters, full knowledge of the classroom context, for example regarding the background of the performance of specific students (Goe et al., 2008). Although teacher surveys can stimulate teacher reflection, underperforming teachers might not have the metacognitive competence to recognize their professional skills and high performing teachers might underestimate these skills (Kruger & Dunning, 1999).

Another way of measuring teaching practices is by means of instructional collections and artifacts. An instructional collection, also called a portfolio, is a collection of materials that is compiled by teachers to provide evidence of their fulfillment of predetermined standards (Goe et al., 2008). Examples of such evidence are lesson plans, assignments, reflective writings, and samples of student work (Gitomer & Bell, 2010). An artifact protocol is a much narrower type of instructional collection, and can for example be focused on the quality of the student assignments that teachers provide. Building an instructional collection can stimulate teacher reflection and can help them improve. It provides insight into the learning opportunities for students on a day-to-day basis. However, it is a time-consuming enterprise for both teachers and assessors, and one might question to what extent teachers’ exemplary work reflects their everyday classroom activities.

Another teaching quality measure, in which not the teaching process but the outcomes of teaching are evaluated, is a measure of a teacher’s added value (Goe et al., 2008). Since schools and teachers teach different student populations, it would be unfair to compare solely the performance of the students across teachers. In value-added models, students’ prior educational attainment, their background characteristics, and the school composition are often taken into account to make comparisons of teachers’ output more fair (Timmermans, Bosker, Doolaard, & de Wolf, 2012). Such measures enable the evaluation of teachers’ contribution to student learning in a cost-effective and non-intrusive way, since most of the required data (test scores) have already been collected for other purposes (Goe et al., 2008). However, the use of value-added models in the context of education is a controversial issue. A review...
of the most recent research on value-added models by Everson (2016) suggests that there are still many questions and concerns that require further research, for example research on how to untangle a teacher's contribution from confounding influences such as classroom materials or specialist support.

1.3 Classroom observation

In this dissertation, the focus is mainly on classroom observation. Although classroom observations have been conducted since the turn of the 20th century or possibly even earlier (Kennedy in Gitomer & Bell, 2010), renewed interest in classroom observation was caused by large scale research projects such as the TIMSS-video study in 1999 (International Association for the Evaluation of Educational Achievement (IEA), 2018) and the Measures of Effective Teaching (MET) project in 2009-2011 (Bill & Melinda Gates Foundation, 2010). During classroom observation, often a classroom observation system (COS) is used. Since COSs are the core topic of this dissertation, the next section first explains what COSs can look like.

1.4 Classroom observation systems

A COS is often thought of as a sheet of paper on which teachers are rated during classroom observation. In this dissertation, the more comprehensive definition of COSs put forth by Bell, Dobbelaer, Klette, & Visscher (2018) is used, in which a COS does not merely encompass scoring tools (scales and items that are scored during classroom observation on a rating scale), but also rating quality procedures (procedures that ensure that raters use the scoring tools accurately and reliably over time, such as rater training or a rater manual) and sampling specifications (a description of the characteristics of the sample of observations including the number of observations that should be conducted per teacher and the length of those observations).

COSs can differ on many aspects (Bell et al., 2018). The scoring tools can focus on different dimensions of teaching. The underlying assumption is that the better teachers score on these teaching dimensions, the better their teaching is presumed to be, and the more students learn. Examples of dimensions that are frequently measured by means of classroom observation systems are classroom management, the explanation of subject matter, cognitive activation (Bell et al., 2018), and student practice (Praetorius and Charalambous, 2018). Which dimensions are presumed to be important to measure, and thus to be included in a COS, is dependent of the
community’s view of teaching and learning (Bell et al., 2018). This community’s view can be located along a continuum that moves from a behaviorist perspective, to a more cognitive perspective, to a more sociocultural view of teaching and learning.

The scoring tools within COSs can either focus on subject specific or generic practices, and can focus on students’ actions, teachers’ actions, or both. Scoring tools can also differ in how discrete/targeted teaching practices are scored during classroom observation, an issue called grain size (Bell et al., 2018). This issue is related to whether a COS is a time-sampling instrument or an event-sampling instrument. In a time-sampling instrument, a rater is asked to count behaviors during a specific time period, for example one minute. In an event-sampling instrument, a rater is asked to score behavior, for example on a 4-point scale from predominantly weak to predominantly strong. The latter is also an example of a high-inference instrument, an instrument that demands much interpretation from the rater; the rater has to decide whether a teacher’s behavior was an example of good quality teaching or not. A time-sampling instrument is an example of a low-inference instrument (van de Grift, 2007).

COSs can also differ in the rating quality measures and the sampling procedures that it provides, in the empirical evidence that is available for the reliability and validity of the COS scores (for example whether raters can score reliable over time), and in where the COS should be placed on a developmental continuum (it takes time to develop a quality COS and to gather empirical evidence for the reliable and valid use of the system; Bell et al., 2018).

This section shows that COSs can differ substantially regarding many aspects. An important question to ask is which quality requirements a COS should meet to be a sound instrument for the measurement of teaching quality. This question has been answered in the first study of this dissertation.

1.5 A framework in support of the development, selection, and use of COSs

Recent publications show that generating valid and reliable scores by means of a COS is not self-evident (e.g., Bell et al., 2012; Cohen & Goldhaber, 2016; Hill, Charalambous, & Kraft, 2012; Hill, Charalambous, et al., 2012; Nava et al., 2018; Sandilos, 2012; van der Lans, van de Grift, van Veen, & Fokkens-Bruinsma, 2016). Many issues need to
be taken into account, for example, raters need to be properly trained, and multiple observations need to be conducted by multiple raters to obtain valid and reliable scores. However, these issues are often not (fully) fulfilled in practice, bringing the reliability and validity of COS scores into question (Coe, 2014). For example, many American states adopted teacher evaluation systems in which only a small number of observations per teacher per year were conducted (Hill and Grossman, 2013). The same happened when Dutch school boards started working with teacher evaluation systems in which a single lesson observation was used for making personnel decisions. Many international researchers also used a single observation for measuring teaching quality after an intervention. A reason for this gap between what we know from research about what classroom observation(s) (systems) require on the one hand and classroom observation practice on the other hand might be that current classroom observation practices feels intuitively valid to many people: “I am a good teacher; I’ll know a good lesson when I see one” (Coe, 2014, p.1).

In order to make classroom observation practices more evidence-based, COSs should be developed based on up-to-date knowledge, be evaluated using critical research standards, and be designed to adequately equip raters and to provide robust scoring designs to users (as suggested by Coe, 2014; Hill et al., 2012). Since there was no framework that brings together the important issues that need to be taken into account when developing, selecting, or using a COS, we answered the research question: Which quality requirements should a classroom observation system meet?

To answer this research question, we developed a COS evaluation framework based on three strands of literature: the literature on COSs, the literature on testing and performance assessment (the Standards for Educational and Psychological Testing, AERA, APA, & NCME, 1999; the COTAN criteria for test quality, Evers, Lucassen, Meijer, & Sijtsma, 2010), and the literature on the argument-based approach to validity (Kane, 2006; 2013). The framework includes quality indicators of a COS such as indicators of its potential for reliable and valid use.

The framework supports COS developers by specifying the topics that need to be considered during the COS development phase and for which topics they need to collect evidence when using the COS. It supports potential COS users by pointing out all relevant topics to consider when choosing between multiple COSs. Finally, through the argument-based approach to validity it assists COS users in evaluating the quality of their COS, as well as in evaluating the evidence available for the reliable use. 
and valid use of the COS in their own context (Kane, 2006, 2013). Although it is important for COS users to evaluate the reliability and validity of the COS scores in their own context, COS developers have the primary responsibility for obtaining and reporting reliability and validity evidence, since potential users need this information to make an informed choice among alternative COSs, and will generally be unable to conduct studies into this prior to the use of a COS (AERA, APA, & NCME, 1999). It is our hope and intention that the framework contributes to the more deliberate design and use of COSs, as well as an increased awareness of the complexity of conducting classroom observations among practitioners, researchers, and local governments (which hopefully leads to the improved use of COSs in the practice of schools). We have applied the framework in the second study of this dissertation, which is described in the next section.

1.6 The quality of COSs

All over the world, many different COSs have been developed by researchers, practitioners, local governments, and commercial parties. The extent to which these COSs meet the quality requirements as included in our evaluation framework had never been reviewed in a systematic way. We therefore conducted this review and answered the following research questions: Which COSs have been developed to measure teaching quality in primary education, what is the quality of the COS materials, and what evidence is available regarding the reliability and validity of the scores these COSs produce?

To answer this question, we conducted a worldwide literature search for COSs in English or Dutch that had been developed for use in primary education. Since our research question also concerns the evidence available of the reliability and validity of the COS scores, an important inclusion criteria was that research into the COSs had to be available. In total, 185 COSs were found, but only 27 of them met all the inclusion criteria. These COSs were reviewed by two reviewers using our COS quality evaluation framework.

The development of the evaluation framework and the review into the quality of the COSs made it, again, very clear how complex it is to obtain valid scores of teaching quality by means of classroom observation. This notion led to the last study in this dissertation, in which raters’ scores are compared to two more cost-effective alternatives for measuring teaching quality.
1.7 Classroom observation scores compared to students’ and teachers’ perceptions of teaching quality

By the time we finished the review, we had the opportunity to start a new research project in which we developed a smartphone application for secondary school students that enabled them to provide feedback to teachers on how they perceive the quality of teaching at the end of a lesson, called the Impact! tool (or IMPACT! app at the time). This tool is appealing in many ways. Teachers can use it at whenever they want to, as much as they want to, and they receive the student feedback right after the lesson in a web-tool. This teaching quality measure is also more cost-effective and less intrusive than classroom observation.

We were interested in how such student perceptions of teaching quality would relate to external raters’ ratings of teaching quality and how these two would relate to teachers’ own perceptions, something that had not yet been researched thoroughly in a single study. We therefore answered the following research question: How do raters’ scores of teaching quality relate to students’ and teachers’ perceptions of teaching quality?

To answer this question we collected data from 25 teachers who participated in a study into the Impact! tool. For three lessons per teacher, we collected the ratings of three external raters based on videotapes of the lessons. Student perceptions were collected through the Impact! tool and teacher perceptions by means of a questionnaire.

1.8 Structure of the dissertation

Each of the three previously introduced studies are reported in a separate chapter. In chapter 2, the framework designed to support the development, selection, and use of COSs is presented. The review of the quality of COSs follows in chapter 3. Chapter 4 presents the findings of the research into the relation between external raters’ ratings of teaching quality and students’ and teachers’ perceptions of teaching quality. Finally, chapter 5 provides a summary and reflection of the main findings of the three studies, as well as some general conclusions and implications. The dissertation concludes with recommendations for future research.

Chapters 2, 3 and 4 are based on three separate research papers, which were all submitted to scientific journals. Each chapter can be read independently, however the chapters may overlap slightly in their theoretical framework.
A framework in support of the development, selection, and use of classroom observation systems
Abstract

Obtaining valid scores using a classroom observation system (COS) is important but not self-explanatory. As with any other quality assessment, various validity issues play a role in the execution of classroom observations and the interpretation of outcomes. A framework that brings together the issues to be taken into account when developing, selecting, or using a COS is lacking, such a framework is therefore presented in this article. The first part of the framework was designed to evaluate the extent to which the design of a COS will allow it to meet the various criteria set by users. The second and third parts of the framework were designed to be used for assessing the reliability and validity of a COS in a specific context.

This chapter is a modified version of the manuscript:
2.1 Introduction

The evaluation of teaching quality can serve various formative and summative purposes. For instance, it may be used to provide feedback to teachers, to determine the focus of professional development at the school level, to make personnel decisions, and to evaluate the quality of teaching in internal or external school evaluations. In the end, these purposes all target the same goal: improving student learning (Archer, Kerr, & Pianta, 2014).

Teaching quality is often measured by means of classroom observation, which can provide meaningful feedback to teachers (Cohen & Goldhaber, 2016) and ensure that teacher training is focused on the teaching processes taking place in the classroom (Lasagabaster & Sierra, 2011). Classroom observation is often considered the most objective tool for measuring teaching practices (Lasagabaster & Sierra, 2011). During classroom observation, often a classroom observation system (COS) is used, which consists of an observation protocol that includes dimensions of teaching on which teachers are scored during observations using a numeric scale. A COS also pertains to the information and activities required to use the protocol as intended. Therefore, a COS also includes rating quality measures (such as a user manual, scoring rules, and rater training) and sampling specifications (for example, specifications regarding the duration and timing of the observations; (Bell et al., 2012; Bell et al., 2018) Although the use of COSs is not new, the Trends in International Mathematics and Science (TIMMS) video study (International Association for the Evaluation of Educational Achievement (IEA), 2018) and the Measures of Effective Teaching (MET) project (Bill & Melinda Gates Foundation, 2010) have recently generated renewed interest in reliability and validity issues surrounding COSs and other teacher evaluation approaches (Bell et al., 2018).

Generating valid and reliable scores when using a COS is not guaranteed. Many authors (e.g., Bell et al., 2012; Cohen and Goldhaber, 2016; Hill, Charalambous, Blazer, et al., 2012; Hill, Charalambous, & Kraft, 2012; Nava et al., 2018; Sandilos, 2012; van der Lans, van de Grift, van Veen, & Fokkens-Bruinsma, 2016) point to issues that need to be taken into account when developing and/or using a COS in order to generate valid and reliable scores. For example, issues regarding the number of raters and observations required for obtaining reliable scores. However, these issues are often not (fully) addressed by COS users or developers, bringing the reliability and validity of the observation scores into question (Coe, 2014). Coe (2014) suggests
that critical research standards and up-to-date knowledge should be applied to the process of developing, implementing, and validating COSs.

As a framework that brings together the important issues to take into account when developing, selecting or using a COS is non-existent, such a framework is presented in this article. These issues are not new but integrated and specified for the use and design of COSs if necessary. For the construction of our framework, we drew from three strands of literature: the literature on COSs, the literature on testing and performance assessment (the Standards for Educational and Psychological Testing, AERA, APA & NCME, 1999\(^2\); the COTAN criteria for test quality, Evers, Lucassen, Meijer & Sijtsma, 2010\(^3\)), and the literature on the argument-based approach to validity (Kane, 2006; 2013). The integration of these three resources has resulted in a comprehensive framework that specifies the topics on which instrument developers need to collect evidence for when designing a COS (e.g., the number of observations and raters required for using their COS). COS developers “must go beyond simply writing instruments: they must create observation systems in which quality observation instruments, well-trained raters, and robust scoring designs are combined to produces reliable teachers scores” (Hill, Charalambous, Kraft, 2012, p.56). The framework supports COS-users in evaluating the quality of a COS-design (e.g., its manual and scoring rubrics) as well as in evaluating the evidence available for the reliable and valid use of a specific COS in their context. It is our intention that the framework contributes to an increased awareness of the complexity of conducting classroom observations among COS developers and users, and to the deliberate design and use of COSs.

The evaluation framework includes three parts. The first part is aimed at evaluating the characteristics of the COS, allowing prospective users to evaluate whether the proposed COS complies with their own goal(s) for conducting classroom observations. The criteria in this part are also meant for the assessment of the theoretical basis of the instrument, as well as the completeness of the guidelines provided to standardize observations and achieve inter-rater reliability. Reliable and valid COS use should be proven empirically. Therefore, in the second and third part of the evaluation framework, the evidence for the reliable and valid use of a COS in a specific context is evaluated.
2.2 The evaluation framework

2.2.1 Part A. Characteristics of the COS

Proposed use
Information in the COS should enable potential users to judge whether it is suitable for their purposes. This requires a description of the constructs the COS aims to measure (such as ‘instruction quality’), who (like teachers in primary education) and/or what (e.g., mathematics lessons) can be observed with the system, as well as a description of the intended (formative and/or summative) use of the observations (Evers et al., 2010).

Theoretical basis
A COS should have a solid scientific basis (AERA, APA, & NCME, 1999); for example, the constructs should have an empirical relation with student learning. The theoretical basis for the constructs measured by means of the COS should be made explicit to potential users by the COS development team. The items in the observation protocol should cover the theoretical construct, thus, the operationalization of the constructs into items is important (Evers et al., 2010).

Quality of the items
Literature on test- and survey construction provides multiple guidelines for formulating items. Incorrectly formulated items can affect the valid use of the observation protocol. Therefore, items should be formulated correctly and not unnecessarily difficult. COS developers should avoid using items that contain double negatives and/or measure multiple aspects simultaneously (Erkens & Moelands, 1992; Evers et al., 2010).

Standardization of observations
Generating completely objective scores using a classroom observation system is impossible, especially with high-inference instruments (i.e., instruments that involve a greater degree of interpretation by the rater). Even the slightest change in the characteristics of the observation can have a substantial impact on reliability. Therefore, COS developers should provide guidelines in order to generate observations that vary minimally between raters.

The number of observations that should be conducted and the number of required raters should be specified in the COS. Hill, Charalambous, and Kraft (2012) showed
that the reliability of the item scores on the Mathematical Quality of Instruction (MQI) observation instrument depends on the number of raters and observations. Even when the same rater scored four lessons of a teacher, a reliability score of .70 (which is often considered to be a minimum score for low-stakes decisions) could not be established. Reliability scores improved by adding multiple raters and observations. The latter was also the case in research by Nava and colleagues (2018) with the ICOR-Math and ICOR-Science observation systems.

Observation scores can be affected by the timing of the observation during the day, week, and year (Casabianca, Lockwood, & McCaffrey, 2015; Pianta, & Hamre, 2009; Sandilos, 2012). Therefore, guidelines concerning the timing of the observation should be specified in the COS, depending on the goal of the COS use. Users could aim to include observations at different times during the week and year (for example, if the COS is used to generate a score about general teaching quality) or users could strive to time the observations similarly to avoid timing influences (for example, if researchers investigate the effect of an intervention).

COS developers should provide guidelines on which lessons to observe, since observations of lessons with a different content domain can result in arbitrary differences in scores (Grossman, Cohen, & Brown, 2014). COS developers should consider the sample of observations used to compute an observation score, depending on the goal of the observation. If observations are used to generate a score about teaching quality in general, the instrument developers should think about the different types of observations needed for obtaining a valid picture of general teaching quality.

When raters follow different procedures during an observation, this could result in different observation scores and thus lower inter-rater agreement. COS developers should therefore provide guidelines for conducting observations like the recommended observation period (e.g., 15 minutes or an entire lesson), when to score the observation form (e.g., at the end of the lesson), how to compute an observed score (e.g., the mean of multiple observations), and how to act during the observation (e.g., if the rater is allowed to walk through the classroom or talk with the students). Hill, Charalambous, Blazar et al. (2012) showed that even the slightest change in observation procedures can have a great impact on the reliability scores. While two school leaders could establish a sufficient level of reliability on a specific item on the MQI, after both observing two entire lessons of a teacher, they would both need to observe four lessons if they would only observe 30 minutes per lesson.
Mashburn, Meyer, Allen, & Pianta (2014) have also suggested that the reliability and validity of scores can be impacted by operational procedures related to length and order of presentation.

**Measures for inter-rater reliability**

Besides guidelines for raters, other measures can be taken to minimize differences in scores between raters and increase inter-rater reliability. These measures are always important, also if the COS is used for formative assessment, because unreliable feedback is a poor basis for improving teaching practice (van der Lans, 2017). Also, when a teacher perceives an observed score as too dependent of the rater, the teacher might not value the feedback and the feedback may not have the intended effect.

An important measure to establish sufficient inter-rater reliability is the development of a rater manual (AERA, APA, & NCME, 1999; Evers et al., 2010), which can include item descriptions, scoring rubrics at the item level that indicate when to assign specific scores, scoring rules to compute an observed score (e.g., the mean of multiple observations), and observation guidelines (e.g., how to act during an observation). An example of such a manual is the K-3 CLASS manual (Pianta, La Paro, & Hamre, 2008).

Another important measure to limit rater influence is rater training (AERA, APA, & NCME, 1999; Hill, Charalambous, Blazer et al., 2012), including for example a discussion of the items and practicing observations using video. The amount of training needed to establish sufficient interrater reliability is dependent on the classroom observation protocol, the extensiveness of the scoring rules, the goal of the observation, and the expertise and experience of the raters. Instrument developers can set an inter-rater reliability threshold that raters should meet before they can use the COS, with or without a certificate (AERA, APA, & NCME, 1999). Even if raters meet the standards after initial training, they can start scoring differently after a while (the so-called ‘rater drift’), for instance, based on new experiences. Casabianca et al., 2015 found that raters initially gave relatively high scores when they started observing, but adjusted their scoring rapidly downward. Rater drift should be assessed (AERA, APA, & NCME, 1999) and could be remedied by providing follow-up training if it occurs (Doabler et al., 2015).

Even with the above measures, it is impossible to establish perfect inter-rater agreement. COS developers could therefore pay attention to how to deal with rater
variance in, for example, the manual.

2.2.2 Part B. Reliable use of the classroom observation system

In part B of this framework, the focus lies on the evidence for the reliable use of an observation system. COS developers have the primary responsibility for obtaining and reporting reliability evidence since prospective users will generally be unable to conduct reliability studies prior to the operational use of a COS, but they need the information to make an informed choice among alternative COSs or other measurement approaches (AERA, APA, & NCME, 1999). Such evidence can indicate that in a specific context (the context for which data are available), the use of the COS can result in reliable scores. Because the scores are very dependent on the context in which they were collected (e.g., the raters, lessons, timing), this information can only be an indication of the reliability for potential users.

Availability of reliability information

Part B of the evaluation framework can only be used if reliability information (of some kind) is available. Different types of reliability can be reported, such as inter-rater reliability, reliability based on inter-item relations, reliability estimates based on item-response theory (IRT), or generalizability theory (Evers et al., 2010). Since COSs are used by raters, information about inter-rater reliability is inherently relevant. The same applies to reliability information about the items in the observation rubric (AERA, APA, & NCME, 1999).

Evaluation of the reliability information

Whether the reliability information can be evaluated as sufficient depends on the intended use of the COS. If the scores are used for high-stakes decisions such as personnel decisions, reliability scores should be at least .80. For decisions with lower stakes (like teacher professionalization activities), a reliability score of at least .70 would be sufficient (Evers et al., 2010).

Evaluation of the research

When evaluating reliability information, the quality of the research generating the information is important. The (analysis) procedures followed in the research should be correct and instrument developers should provide enough information to enable the thorough judgement of the reliability of the COS scores (Evers et al., 2010). Hallgren (2012) describes common mistakes that researchers make in assessing and reporting inter-rater reliability. Many researchers report percentages of agreement,
while this is rejected as an adequate measure of inter-rater reliability. The percentage of agreement does not correct for agreement that would be expected by chance and therefore overestimates the level of agreement. Other measures are available that do account for chance agreement, such as Cohen’s kappa and intra-class correlation. Many variants are available and, according to Hallgren (2012), many researchers do not report which statistic or variant was used in inter-rater reliability analyses and/or fail to use the correct statistic. Factors such as the metric in which a variable was coded (e.g., nominal or interval) and the design of the study (for example, whether subjects are rated by all raters or only by a subset of the raters) must be considered when the most appropriate statistical test is selected (Hallgren, 2012).

In the reliability research, the system should be used as proposed in the manual (Evers et al., 2010). The samples used should match the target group of the COS (e.g., a COS designed for primary education should be implemented in primary education). A clear description of the samples helps potential users judge the extent to which the reported data applies to their own population (AERA, APA, & NCME, 1999).

2.2.3 Part C. Valid use of the classroom observation system
In this third part of the evaluation framework, the focus is on the evidence for valid use of the COS using the argument-based approach to validity (Kane, 2006; 2013). In this approach, the network of inferences and assumptions leading from the sample of observations to the conclusions and decisions based on the observations is specified (the interpretive argument) and evaluated (the validity argument). By outlining the interpretive argument, the reasoning behind the proposed interpretations and uses of the observations becomes explicit, allowing for their understanding and evaluation (Kane, 2006). Four common inferences in interpretive arguments for the use of COSs are the scoring inference, the generalization inference, the extrapolation inference, and the implication inference, each of which will be elaborated upon below.

According to Kane (2006), each inference within the interpretive argument could be seen as a practical argument as described by Toulmin (1958), see Figure 2.1. In each inference, a claim is made based on the data. In the scoring inference for example, an observed score (claim) is generated based on a sample of observations (data). Warrants (a rule or a principle) provide support for the legitimacy of the relation between the data and the claim. A warrant for the scoring inference is that measures were taken to score accurately and consistently. However, warrants generally are not self-evident and have to be justified by evidence called backing. An example of
backing for the above warrant is the training of raters. Toulmin (1958) describes two additional components in his analysis of inferences: a qualifier (the strength of the claim) and conditions of rebuttal (conditions under which the warrant would not apply). The claim in each inference serves as the data for the next inference.

![Figure 2.1 Analysis of inferences (Toulmin, 1958)](image)

In the validity argument, the warrants and backing for each inference in the interpretive argument are reviewed critically in order to evaluate the validity of the claim. Which warrants and backing are important to review depends on the proposed interpretation (Kane, 2006), which will vary for each COS and even for different uses of the same COS. For example, different forms of backing for rater bias is needed when teachers are observed by familiar raters instead of external raters. In this part of the evaluation framework, warrants and backing that are often relevant for evaluating valid use of a COS are provided.

Interpretive arguments are often not explicitly specified by instrument developers. Based on the information that is available about the COS, potential users can specify the underlying interpretive argument and subsequently evaluate the evidence provided by the instrument developers for the valid use of the COS within the research context. This cannot simply be generalized to the context and use of potential/other users (Hill, Charalambous, Blazer et al., 2012). Therefore, users of a COS should also specify their own interpretive argument and gather as much evidence for the validity argument as possible. The evaluation framework can therefore support users in formulating their interpretive argument and in making well thought out choices about the implementation of the COS in their own context and for their own purpose.
In the following sections, the four possible inferences that can make up an interpretative argument for COSs (see Figure 2.2) and the associated warrants and backing that can be used for the validity argument are described. Not all four inferences are necessarily always relevant for evaluation. This is dependent on the use of the COS. For example, if users of a COS do not have the intention to generalize the observed score, the generalization inference will be irrelevant.

Figure 2.2 Possible interpretive argument (based on Kane, 2006; Wools, Eggen, & Sanders, 2010)

Scoring inference

The scoring inference connects a sample of observations to an observed score. The following three warrants can support this inference:

» The scoring rules are substantiated (empirically)

The ‘scoring rules’ in the evaluation framework refer to: (1) scoring rules at the item level that help a rater distinguish between scores on a scale, and/or (2) scoring rules to compute an observed score (when multiple observations are conducted). The scoring rules should be appropriate, meaning that “the scoring rules define interactions in ways and at levels that are suitable for the classroom interactions they purport to measure” (Bell et al., 2012, p. 66). Possible backing for the scoring rules could be that the observation protocol has been explicitly developed on the basis of solid theory, research or standards, or that the rules are supported by relevant stakeholders like experts in the field (Bell et al., 2012; Kane, 2006).

Backing for the scoring rule can also be obtained empirically (Kane, 2006). If it is expected that performance varies across the full scoring range and if raters only rate one or two points on the scale, the scoring rule might not be appropriate (Bell et al.,
2012). Statistical analysis can also clarify if the statistical models used (e.g., scaling) are appropriate (Kane, 2006) and can provide information about the psychometric quality of the items (Evers et al., 2010).

» **Measures were taken to score accurately and consistently**

Several measures can be taken to ensure that raters score accurately (as intended) and consistently as much as possible, such as scoring rules, rater training, and a reliability threshold for raters (as described in part A of this framework). Research by Hill, Charalambous, Blazer, et al. (2012) suggests that a rigorous selection of raters based on their inter-rater reliability increases the likelihood that the raters will understand and use the items as intended by the instrument developers.

Empirical evidence for inter-rater reliability can show that a COS can be used accurately in a specific context (Kane, 2006). Research can also show that raters can use the system consistently over time (Bell et al., 2012).

» **Attention is paid to rating bias**

Observation scores could be biased. A main source of bias comes from the ways raters are assigned to lessons. If a teacher is always observed by the same (familiar) rater who scores somewhat differently from other raters, then the scores will be biased. Another important source of bias comes from the ways in which raters assign scores. Raters could rate a specific group of teachers/students or types of classrooms differently because of their own education/expertise, beliefs and prejudices (Bell et al., 2012). Bias can be reduced by working with multiple raters and by training raters.

Statistical analysis, such as Differential Item Functioning (DIF) analyses (AERA, APA, & NCME, 1999), can provide insight into the extent to which scores are biased.

**Generalization inference**

Usually it is not the goal to make claims about teacher performance during just a particular lesson. Generally, the goal is to make claims about teacher performance within some larger domain of tasks over some range of occasions and conditions of observations, called the universe of generalization (Kane, 2013), see Figure 2.2. If a teacher is, for example, observed for personnel decisions, one might want to generalize the observed score to all the lessons the teacher taught that year. In the generalization inference, the interpretation of the observed score is expanded from a claim about a specific set of observations to a claim about expected performance
A framework in support of the development, selection, and use of COSs

across the universe of generalization, a universe score (Kane, 2006).

Two important preconditions for generalization are (1) a representative sample of observations from the universe of generalization, and (2) a large enough sample to control for sampling error. Whether a sample is representative or not is more often based on logic than statistics, because statistical sampling assumptions are rarely met (Kane, 2006). According to Kane (1996), it would be reasonable to assume that the sample is representative “when a serious effort has been made to draw a representative sample from the universe of generalization and there is no indication that this effort has failed” (Kane, 2006, p.35).

The warrant for this generalization inference in the evaluation framework focuses on the extent to which COS developers are explicit about generalization. Based on the information they provide, a judgement can be made of whether the proposed generalization is plausible and feasible for prospective users. Hill, Charalambous, Blazer, et al. (2012) suggest that reliable estimates of the construct should be achievable under reasonable budget constraints, which seems to be an important consideration for potential users of a COS. If users of a COS implement the system differently than implied by the instrument developers, this could have significant consequences for the generalizability assumption.

» In the COS, information about generalization possibilities is provided explicitly

As explained, characteristics of the observed lesson, such as the length of the observation, the timing of the observation, and the content domain of the lessons, can affect the observed score. According to Bell et al. (2012), it is important to examine the effect of these factors on the observed score and to take these factors into account by adjusting scores, or by using sampling schemes. As already indicated, it is also important to take the sample size into account to control for random error (Kane, 2006). Therefore, the number of observations required for generalization should be elaborated in a COS.

Empirical evidence for the generalization inference could be collected in a generalization study or a reliability study (Kane, 2006).

*Extrapolation inference*

The universe of generalization is often a (very) small subset of what is really of interest (the target domain), see Figure 2.2. For example, if a COS aims to measure teaching
quality, only a small subset of what teaching quality involves is measured by means of the COS. However, the fact that some issues of the construct are hard to measure (by means of observation) does not mean that they are not part of the target domain (Kane, 2006).

In the extrapolation inference, the interpretation of the observed score is extended from the universe of generalization to the target domain, yielding the target score. The score does not change; rather its interpretation is extended. A solid evaluation of the extrapolation inference requires both analytical and empirical evidence (Kane, 2006). To evaluate the extrapolation inference, the extent to which the universe of generalization is related to the target domain is included in the evaluation framework.

» The score on the universe of generalization is related to the target domain

“Extrapolation depends, at least in part, on the relationship between the universe of generalization and the target domain” (Kane, 2006, p. 35). This relationship is more plausible if the universe of generalization covers a large part of the target domain. The more the universe of generalization covers the target domain, the more plausible the extrapolation inference. For instance, extrapolation of an observed score to the domain of teaching quality is more plausible if the COS measures many aspects of teaching quality. However, a broad universe of generalization can have implications for the generalization inference, because it is more difficult to draw a sample of observations that is representative for a broad universe of generalization. So there is a trade-off and instrument developers should try to find a compromise that supports both generalization and extrapolation (Kane, 2013). Extrapolation is also more plausible if the processes during the measurement are the same as in the target domain (Kane, 2013), which applies to the use of a COS since it is used to observe ‘real’ lessons.

Empirical support for the validity of the extrapolation could be obtained by examining the relations between the observed scores and scores obtained from other measures (Kane, 2006). Since a criterion score for the target domain of COSs is probably lacking, the observed score could be correlated with other measures of the target domain such as student perceptions of teaching quality, other COSs, or student achievement.

**Implication inference**

The implication inference connects the target score with the interpretation and use of the score (Bell et al., 2012). Is the interpretation and use of the scores as proposed by
the development team appropriate? The more ambitious the interpretation, the more backing is needed to support the claim (Kane, 2013). Evidence for this implication can be analytical and empirical. This leads to two warrants:

» The proposed use and implications of the score seem appropriate based on the theoretical construct

In order to make valid claims regarding the observed score, the claims should at least match with the target domain and the used classroom observation instrument (Kane, 2006).

» The proposed use and implications of the score seem appropriate based on empirical evidence

Specific implications of the interpretation and use of the score could be checked empirically (Kane, 2006). When the COS is used to provide feedback to teachers, it could be explored whether the feedback led to improvement in the areas indicated for improvement by the observation system (Bell et al., 2012). If observation scores are used for personnel decisions, one might expect the scores to be more or less stable, which could be checked empirically. The same applies to scores of different groups of teachers. One might expect, for example, beginning teachers to score lower than experienced teachers.

2.3 Discussion

The purpose of this article is to present an evaluation framework that can be used as a reference by developers and users of COSs. It is our hope that this will strengthen general awareness of the complexity of using COSs, and support the deliberate design and use of COSs.

The framework is demanding in terms of the requirements for COSs’ use and we are aware that it will be hard to meet all the presented standards completely. We believe, however, that it is important for COS-users to develop their own interpretive argument and to decide which evidence is most important for the reliable and valid use of the COS in their specific situation. As users of a COS in many instances will be limited in their possibilities to collect evidence about COSs (before they start using a COS), instrument developers, in our view, have an important role to play in providing comprehensive systems for classroom observation that have been researched extensively. It is their task to present the research findings to potential users in an accessible and clear manner (for example, in the user manual).
Chapter 2

As our framework shows that generating valid and reliable indicators of teaching quality is very complex, we recommend that users of COSs are careful in drawing conclusions based on classroom observations, also if the observations are conducted for formative evaluation.

Finally, we used the framework to review the qualities of the worldwide available COSs for the measurement of teaching quality; the results are presented in chapter 3.

Footnotes
1. The 2014 edition of the Standards for Educational and Psychological Testing was not available yet during the development of the framework. The areas that received particular attention in the 2014 revision are not relevant for our framework. The information relevant for the development of the framework does not differ substantially between both versions.
2. The COTAN criteria are used by the Dutch Committee on Tests and Testing (COTAN) to assess the quality of psychological tests available in the Netherlands. COTAN has audited over 750 tests published for professional use.
A framework in support of the development, selection, and use of COSs
3

The quality of classroom observation systems for measuring teaching quality in primary education: a systematic review
Abstract

Teaching quality is often measured by means of classroom observation using a classroom observation system (COS). Generating valid and reliable scores when using a COS is not self-explanatory. Many authors point to issues that need to be taken into account when developing, selecting, and using a COS in order to generate valid and reliable scores. A framework that brings together these issues was developed and followed by an extensive worldwide literature search for COSs, resulting in 27 COSs that met the inclusion criteria. All COSs were reviewed by two reviewers. Reviewers were, on average, positive about the scoring tools, however, insufficient empirical evidence was found for the reliable and valid use of the scores for most COSs. More research into COSs is needed to determine how reliable and valid scores can be provided to teachers who (want to) use a COS.

This chapter is a modified version of the manuscript: Dobbelaer, M. J., & Visscher, A. J. (submitted). The quality of classroom observation systems for measuring teaching quality in primary education – A systematic review.
3.1 Introduction

Classroom observation is widely used by researchers and practitioners in the field of education to rate/evaluate teaching quality for research purposes or for the formative/summative evaluation of teachers. Often, a classroom observation system (COS) is used for such observations. In the definition by (Bell et al., 2018), a COS comprises three aspects: scoring tools, rating quality procedures, and sampling specifications. The scoring tools consist of the actual scales and items that are scored during classroom observation and are scored on a rating scale. Rating quality procedures, such as rater training or a rater manual, refer to the procedures in the system to ensure that raters use the scoring tools accurately and reliably over time. The sampling specifications in a COS consist of a description of the characteristics of the sample of observations (e.g., the number of observations that should be conducted per teacher and the length of those observations) in order to allow for the generalization of the observed score to a broader context (e.g., the quality of teaching during a specific period of time).

Over the years, many different COSs have been developed and others are still in development. For a new research project, a COS was needed to measure teaching quality in primary education. In an attempt to obtain a quick overview of which COSs had already been developed, it became clear that it is difficult to obtain such an overview as a potential COS user and even more difficult to evaluate the quality of the existing COSs. Therefore, this review was conducted to answer the following questions: Which COSs have been developed to measure teaching quality in primary education, what is the quality of the COS materials and, what evidence is available regarding the reliability and validity of the scores these COSs produce?

3.2 The evaluation framework

A selection of COSs were reviewed using an evaluation framework consisting of three parts, evaluating the COS materials, the evidence for the reliability of the COS scores, and the evidence for valid use of the COS scores. The criteria in the framework were drawn from three strands of literature: the literature on COSs, the literature on testing and performance assessment (the Standards for Educational and Psychological Testing, AERA, APA & NCME, 1999; the COTAN criteria for test quality, Evers, Lucassen, Meijer, & Sijtsma, 2010), and the literature on the argument-based approach to validity (Kane, 2006; 2013). The evaluation framework and the underlying theory is outlined in the following paragraphs.
3.2.1 The COS materials

Scoring tools

Each COS in this review includes a scoring tool consisting of items (also named criteria, dimensions, elements or indicators) that are scored on a rating scale. All COSs measure the quality of teaching, however, they can focus on different dimensions. Bell et al. (2018) describe nine dimensions of teaching that are fundamental to students’ learning and development that a COS can focus on: a safe and stimulating classroom climate, classroom management, the involvement and motivation of students, explanation of subject matter, the quality of subject matter representation, cognitive activation, assessment for learning, differentiated instruction, and teaching learning strategies and student self-regulation. The COS items can be derived from different theories, research, or standards, but all should have a solid scientific basis (AERA, APA, & NCME, 1999) and the items should cover the theoretical constructs (Evers et al., 2010). Although the theoretical basis of the COSs was evaluated in this review, it was not feasible to also evaluate the quality of the research underlying the COSs.

The items in the COSs can be subject specific (e.g., designed to capture the quality of mathematic teaching) or generic (items that can be used across subjects), and can focus on teachers’ actions, students’ actions, or both (Bell et al., 2018). The number of items included in the scoring tools can differ as well as the response categories in the rating scale. Strong (2011) points out that a large number of items can be problematic for raters because “there is an upper limit of an rater’s ability to match his or her responses to a given set of stimuli” (the channel capacity; Strong, 2011, p.88). Although utilizing a small number of items may reduce an rater’s cognitive load and be sufficient for evaluating teaching quality well, more items enable the provision of richer feedback to teachers on their strengths and weaknesses, which is needed for improvement (Marzano, 2012).

The literature on test and questionnaire construction provides many requirements for formulating items. In this review, criteria (based on Erkens & Moelands, 1992; Moelands, Noijons, & Rem, 1992) were used to evaluate the quality of the items in the scoring tools. These quality criteria relate to whether the items in the COSs (a) are grammatically correct, (b) do not include complicated linguistic constructions, words that can have a different meaning when the emphasis is shifted, double negatives, and ambiguities, (c) do not include unnecessarily difficult words, inserts, or negative statements, (d) are formulated clearly to prevent misunderstandings, and (e) measure a single behavior at a time.
A description of the scoring tools should enable potential users to judge whether a COS is suitable for their purposes and should therefore include a description of the constructs a COS aims to measure, the type of use for which the COS has been developed, and who/what can be observed using the system (Evers et al., 2010).

**Rating quality measures**

Classroom observations are conducted by raters and measures can be taken to increase inter-rater reliability (IRR). These measures seem important for both summative and formative teacher evaluation, as unreliable feedback is a poor basis for improvement (van der Lans, 2017).

An important measure for establishing sufficient IRR is the development of a rater manual (AERA, APA, & NCME, 1999; Evers et al., 2010). Such a manual can include item descriptions, scoring rules that indicate when to assign which scores, and scoring rules regarding how to compute an observed score. A manual can also include guidelines concerning the recommended observation period (e.g., 30 minutes), when to score the observation form (e.g., at the end of the lesson), and how to act during the observation (e.g., stay at the periphery of the action), since this is likely to affect the scores provided by raters. Considering that multiple studies have shown that the reliability of the scores improves if the number of raters increases (Hill, Charalambous, & Kraft, 2012; Nava et al., 2018), manuals can also include guidelines concerning the number of raters per teacher.

The training of raters is another important measure to limit rater influence on COS scores (AERA, APA, & NCME, 1999; Hill, Charalambous, Blazer et al., 2012). Instrument developers can set an IRR threshold that raters should meet (after training) before they can use the COS (AERA, APA, & NCME, 1999). Follow-up training can be important too because, after a while, raters can start scoring differently as a result of, for instance, new experiences (the so-called ‘rater drift’; Casabianca, Lockwood, & McCaffrey, 2015; Dobbelaer et al., 2015).

**Sampling specifications**

If the observed scores are generalized to a broader domain than what was observed (for example, teaching quality over a school year), which is often the case, then sampling specifications become important as the observation(s) should be representative of the context to which the score is generalized. Important specifications are the required number of observations per teacher (Hill, Charalambous, & Kraft, 2012; Nava et al.,
2018), the duration of those observations (Hill, Charalambous, Blazar et al., 2012),
the types of lessons (Grossman et al., 2014), and the timing of the observations during
the day, week, and year (Casabianca et al., 2015; Pianta & Hamre, 2009; Sandilos,
2012).

3.2.2 Evidence for reliability
The second part of the evaluation framework concerns the evidence for the reliability
of the scores. Providing reliability evidence is primarily the responsibility of COS
developers, since prospective users need this information to make an informed
choice among alternative COSs or other measurement approaches, and prospective
users will generally be unable to conduct reliability studies prior to the operational
use of a COS (AERA, APA, & NCME, 1999).

Reliability evidence was evaluated in this review if reliability information on both
the scoring tools and the raters was available. In this evaluation, the intended use
of the COSs was taken into account. A higher reliability coefficient (or a similar
measure) is necessary for high-stakes decisions (e.g., tenure decisions) than for
low-stakes decisions such as teacher professionalization activities. The quality of the
research was also taken into account (as suggested by Evers et al., 2010), namely,
whether the (analysis) procedures followed were correct, whether the research had
been conducted in the target group of the COS (e.g., a COS designed for primary
education should be investigated in primary education), and whether developers
provided sufficient information to thoroughly judge the reliability of the COS scores.

3.2.3 Evidence for validity
The third part of the evaluation framework focuses on the evidence for the valid use of
a COS, using the argument-based approach to validity (Kane, 2006). The argument-
based approach specifies (the interpretive argument) and evaluates (the validity
argument) the relationship between the sample of observations and the inferences,
assumptions, conclusions, and decisions based on them.

The interpretive argument that is specified for COS use often includes four inferences:
the scoring inference, the generalization inference, the extrapolation inference, and
the implication inference (Bell et al., 2012; Kane, 2006). In the scoring inference,
an observed score is generated based on a sample of observations. For example, a
composite score is computed based on the observation of five lessons. Often, claims
about teacher performance are made that are more general than what is justified by
the sample of observations. In the generalization inference, the interpretation of the observed score is expanded to a larger domain of tasks over a range of occasions and conditions of observations, i.e., the universe of generalization. For example, the score based on five lessons is generalized to all the lessons of a teacher during a school year. The universe of generalization might be only a subset of what is really of interest. In the extrapolation inference, the interpretation of the score is further expanded to the larger domain of interest, i.e., the target domain. In our example, the observed score can for instance be used to make claims about teaching quality in general. Finally, the score is interpreted in the implication inference. For instance, the teacher in our example may receive a higher salary based on the five observations.

Kane (2006) argued that within the interpretive argument each inference can be considered to be what Toulmin (1958) called a practical argument: in an inference data lead to a claim. For example, in the scoring inference, a sample of observations (data) leads to an observed score (a claim). Rules or principles (warrants) can support the relation between the data and the claim. An example of a warrant for the scoring inference is that measures were taken to score accurately and consistently. The warrants themselves also have to be justified by evidence which is called backing. In the case of our example, a backing for the warrant is that raters are trained.

In order to evaluate the claims made in each inference, the warrants and backing for each inference are evaluated in the validity argument. Which warrants and backing are evaluated depends on the proposed interpretation of scores (Kane, 2006), which can vary across COSs and even across different types of use of the same COS. In this review, the validity evidence for each COS was evaluated with a fixed set of warrants and backing for each inference, based on Bell et al. (2012) and Kane (2006; 2013). These warrants and backing are regarded as always relevant when evaluating the valid use of a COS and are presented in Table 3.7, Table 3.8, and Table 3.9, as well as in Appendix A. For a more detailed description of the warrants and backing, see chapter 2.

3.3 Method

Before searching for COSs, a review protocol (available in Dutch) was developed in collaboration with an information specialist of the University of Twente. This protocol included the aim of the review, the research questions, the inclusion and exclusion criteria, and the search strategy.
3.3.1 Inclusion and exclusion criteria
Given the focus of the review, a COS was only included if it measured (an aspect of) teaching quality and if it had been developed for use in primary education. The COSs in this review had to be useful to a broad range of stakeholders, including teachers, school boards, districts or researchers. Therefore, a COS was excluded if it was meant to be used only by in-service teachers, if the focus was on teaching a subgroup of students (e.g., high-achieving students), teaching outside the classroom, or involved the implementation of a specific intervention or method. Only generic systems or subject-specific systems focusing on language arts, mathematics or science were included. Furthermore, the observation data provided by the COS had to be interpretable by this broad range of stakeholders and for this reason, time sampling observation systems, in which specific teaching behaviors are counted during a number of time samples, were excluded because the data that these COSs provide are often difficult to interpret by for example teachers and school leaders.

Since the main focus of our review concerns the quality of the COSs, research into the reliability and validity of the COS scores had to be available. COSs were included in the review if some empirical information was available regarding the use of the COS in primary education. At this point, no criteria concerning the quality of that research were used.

For practical reasons, only COSs available in English and Dutch were included; due to the availability of the COS materials, only COSs with at least one publication since the 1st of January 1990 were included. Furthermore, to be included, a scoring rubric had to be available.

3.3.2 Search strategy
COSs were searched for using a five-step search strategy based on Littell, Corcoran, and Pillai (2008), including a systematic literature review, contacting experts in the field, a manual search through relevant journals, using the backward-snowball method, and reviewing the grey literature and conference programs. This search was conducted between September 2014 and October 2015.

The systematic literature review was undertaken using four international databases (Educational Resources Information Center (ERIC), Web of Science, Scopus, and PsycINFO), two Dutch databases (Narcis and PiCarta), and one dissertation database (ProQuest Dissertation & Theses Global). In all databases, we searched for
publications that included a search term for classroom observation or classroom observation system in the title, abstract, or subject/topic (observation, evaluation, inspection, supervision, observation measure, observation system, observation protocol, observation instrument, rat* scale, assessment instrument, measure) that appeared within 15 words of a search term for teaching or classroom (teach*, class or lesson). The results were narrowed down using search terms for primary education (e.g., primary education, elementary school, k-12, grade 1), research terms (e.g., valid*, reliab*, alpha, factor analyses, inter-rater), and a restriction regarding publication date (1990-2014). The full electronic search strategy for ERIC is presented in Appendix B. The search resulted in 6,567 publications. Seventy of these publications, in which either a COS was described or the research topic, seemed to meet the inclusion criteria based on the title and abstract, or the full text.

The contacted field experts were all researchers conducting research into teaching evaluation, or had developed or used a COS in their research. A total of 92 researchers from 13 different countries were contacted. They were asked to name COSs that met the most important inclusion criteria. Sixty-three researchers responded and 33 of them named one or more COSs.

A manual search was conducted through 17 international journals (e.g., Educational Assessment, Educational Evaluation, Educational Research, and Language Teaching Research) and one Dutch journal (Pedagogische Studiën). The title and the abstract of publications between 1990 and 2014 (that were available online) were scanned. A manual search through online available programs of the Educational Research Days (a Dutch annual research conference) was also conducted. These manual searches resulted only in a few new COSs that seemed to meet the inclusion criteria based on the title and abstract or the full text.

Backward-snowballing (reviewing the reference list of identified articles) resulted in more relevant COS publications and grey literature. Especially compendia of instruments useful in education (Denham, Ji, & Hamre, 2010; Halle, Whittaker, & Anderson, 2010; Malone et al., 2010; Overdiep, 2014) yielded new COSs that seemed to meet the inclusion criteria.

Eventually, this five-step search strategy resulted in 185 unique COSs that initially seemed to meet our inclusion criteria.
3.3.3 Final pool of COSs

Out of the 185 COSs found in our systemic search, 66 of them did not meet our inclusion criteria based on the information found. For the other 119 COSs, the initial search provided insufficient information to make a judgement. Sometimes our initial search only resulted in the name of the COS. In other cases, the information from the initial search did not include research into the reliability and validity of the COS scores and as the availability of this information elsewhere was unclear, it was uncertain whether the COS could be included. Therefore, an additional systematic search was conducted, using the title of the COS, for each of these 119 COSs in four international databases (ERIC, Web of Science, Scopus, Psychinfo) and Google scholar. If this search also failed to provide enough information to judge whether the COS met our inclusion criteria, the main developer of the COS or author of publications on the COS was contacted for more information.

Eventually, 27 COSs met our inclusion criteria. Table 3.1 shows the reasons for excluding the other 158 COSs from our review. If one of the inclusion criteria was not met, the other criteria were not evaluated anymore. The most common reasons for exclusion was the absence of research into the COSs or the kind of data that the COSs provide (data that is not easy to interpret by a broad range of stakeholders).

<table>
<thead>
<tr>
<th>Reason</th>
<th>Number of COSs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absence of research into psychometric properties</td>
<td>36</td>
</tr>
<tr>
<td>COS not useful for non-researchers</td>
<td>33</td>
</tr>
<tr>
<td>Measuring something other than teaching quality</td>
<td>21</td>
</tr>
<tr>
<td>Measuring teaching quality of subgroups of students</td>
<td>15</td>
</tr>
<tr>
<td>Not developed for or validated in primary education</td>
<td>13</td>
</tr>
<tr>
<td>Too similar to an included instrument</td>
<td>11</td>
</tr>
<tr>
<td>Lack of information</td>
<td>6</td>
</tr>
<tr>
<td>COS not available in English/Dutch</td>
<td>6</td>
</tr>
<tr>
<td>Other reason</td>
<td>17</td>
</tr>
<tr>
<td>Total</td>
<td>158</td>
</tr>
</tbody>
</table>

Table 3.1 Reasons for exclusion of 158 COSs
3.3.4 Review of the COSs

Description of the COSs

A range of sources provided information about the COSs: user manuals, manuals for COS use in research projects, peer reviewed publications, dissertations, grey literature, websites, and personal contact with the authors. All relevant information was described in an overview document for each COS, which included information on the general characteristics of the COS (e.g., the authors and the costs of the COS), the assessment domain (e.g., the items and the underlying theory), the observed score (e.g., the sample of observations and the raters), the proposed use (e.g., the intended conclusions, decisions, and actions based on the observed score), and a description of the research into the COS (e.g., the research design, sample, and results). See Appendix C for a complete list of the topics covered in the overview.

The overview document for each COS was drawn up by the first author between November 2015 and February 2017. During this process, (another) systematic search was conducted for each of these 27 COSs in four international databases (ERIC, Web of Science, Scopus, Psychinfo) and Google scholar, using the title of the COS, in order to further increase the probability that all relevant publications were found. Backward-snowballing was also used to find more publications concerning the COSs. When information in the overview document was missing, the authors of the COSs were contacted for more information. An overview of all the included publications per COS can be found in Appendix D.

Evaluation framework

The 27 COSs were reviewed based on the evaluation framework that was developed for this review. All criteria in the evaluation framework can be found in Appendix A. The evaluation framework consists of 40 to 55 criteria, depending on which items in the evaluation framework were relevant for the evaluation of a particular COS. The reviewers scored all criteria, in an online form, on a dichotomous scale (insufficient or sufficient). The reviewers were instructed to assign the score ‘insufficient’ if there was not enough evidence for a sufficient assessment of a criterion. The reviewers were obliged to motivate every response.

During the review process and analysis, some of the criteria in the evaluation framework turned out to be problematic (e.g., because the scoring rules were not specific enough) or not relevant enough. It was also opted to not report on the evaluation of the implication inference. The criteria that were excluded from the
analyses are indicated in Appendix A with an asterisk.

**Reviewers**

All 27 COSs were reviewed by two reviewers: the first author and one of five second raters. Four of the second raters work at the primary education department of the Dutch Inspectorate of Education as a school inspector. They all possess a Master’s degree in Social Sciences and two of them also have a Doctoral degree in Educational Sciences. The fifth second rater was a researcher at the University of Twente with a Doctorate degree in Educational Sciences. The second raters, depending on their availability for this project, reviewed four to seven COSs.

All raters were trained for the use of the evaluation framework by the first author, through the discussion of the evaluation framework, practicing rating on two COSs, and by discussing the ratings. The agreement between the first author and the second raters was moderate (Cohen's Kappa = .46). Criteria in the evaluation framework that were not used in the analyses were also excluded from this inter-rater reliability analyses.

The results also include an overview of the nine dimensions of teaching (Bell et al., 2018) that are included in the 27 COSs. Both the first author and the second author (not a reviewer) individually indicated for each COS which teaching dimensions were covered in the COS, and, if necessary, the results were discussed to reach a consensus.

**Analysis**

The results of this review are descriptive, consisting of a description of the scoring tools based on both the information in the overview documents and the evaluation framework, descriptive statistics of the results of the review based on the criteria in the evaluation framework, and, if relevant, a description of the reasons for a score. These reasons were analyzed qualitatively by open- and axial coding.

As agreement between the reviewers was moderate, we decided not to report the review results per COS but to provide general results per evaluation framework component.
3.4 Results

The results of our analyses are presented below under the three main components of the theoretical framework: the COS materials, the reliability of the COS scores, and their validity.

3.4.1 The COS materials

The scoring tools

The next section provides an overview and an evaluation of the scoring tools in the COSs by discussing their general characteristics, content, theoretical basis, and the quality of the items.

» General characteristics

Table 3.2 provides an overview of the 27 COSs that met the inclusion criteria and were thus included in the review. The table includes the primary reference and some general characteristics of each COS. An extended description of each COS and all references to the publications that were included in the review can be found in Appendix D. Table 3.2 shows that the COS differ in many respects.

Most of the COSs were developed for general use (any primary school subject). Some were developed specifically for use in language, mathematics, science, or both mathematics and science classes. All COSs were developed for use in primary education, however eight COSs focus on a specific group within primary education (e.g., K-3), and 12 COSs can be used more widely (e.g., K-12). CLASS includes different versions for different age groups. Each CLASS-version includes the same dimensions, allowing for the measurement of teaching quality across grades (however, the elaboration of each dimension is age group specific). Developers do not always clearly specify which teachers (in what grade) and which subject(s) can be observed with their COS. For 13 COSs, at least one of the two reviewers indicated that this was not specified well enough. In those cases, the information in Table 3.2 is based on how the instrument developers used the COS in their research.

All COSs are intended to be used for rating teaching quality. However, the number of items for doing so differs strongly (from 6 to 78) and items are scored on rating scales that range from 2- to 7-point rating scales. In some instruments, the number of items is not fixed but depends on characteristics of the lesson, such as the type (CMI), content (Marzano), subject (DCO, SAMPI, and IQA), or whether technology is used.
<table>
<thead>
<tr>
<th>COS</th>
<th>Full name</th>
<th>Main reference</th>
<th>Grade</th>
<th>Subject</th>
<th>Teacher / student</th>
<th>Scale points</th>
<th>Items</th>
<th>Teacher / student</th>
<th>Purpose</th>
<th>Used by others</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIMS Instrument</td>
<td>Classroom Atmosphere Management System</td>
<td>Roebrig &amp; Christesen, 2010</td>
<td>K-12</td>
<td>G</td>
<td>Both</td>
<td>4/7</td>
<td>75</td>
<td>Both</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>ASOS</td>
<td>Adult Style Observation Schedule</td>
<td>Van Hiedegem, Gadey, Gadey,</td>
<td>K-6</td>
<td>L</td>
<td>Teacher</td>
<td>7</td>
<td>75</td>
<td>Teacher</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>CLASS K-3</td>
<td>Classroom Assessment Scoring System (K-3)</td>
<td>Pianta, La Paro, &amp; Hamre, 2008</td>
<td>K-3</td>
<td>M</td>
<td>Both</td>
<td>10</td>
<td>12</td>
<td>Both</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>CLASS UE</td>
<td>Classroom Assessment Scoring System (Upper Elementary)</td>
<td>Pianta &amp; Hamre, 2012</td>
<td>K-6</td>
<td>M</td>
<td>Both</td>
<td>7</td>
<td>75</td>
<td>Both</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>CMI Protocol</td>
<td>Comprehensive Mathematics Instruction Observation Protocol</td>
<td>Womack, 2011</td>
<td>K-6</td>
<td>M</td>
<td>Teacher</td>
<td>76</td>
<td>78</td>
<td>Teacher</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>CSS-Form</td>
<td>Classroom Strategies Scale-Observer Form</td>
<td>Reddy, Fabiano, Duduk, &amp; Hsu, 2013a</td>
<td>K-6</td>
<td>M</td>
<td>Teacher</td>
<td>44</td>
<td>7</td>
<td>Teacher</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>CTI</td>
<td>Constructivist Teaching Inventory</td>
<td>Greer, 1997</td>
<td>K-6</td>
<td>S</td>
<td>Teacher</td>
<td>44</td>
<td>7</td>
<td>Teacher</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Note: G = General, L = Language, M = Mathematics, S = Science, b R = Research, F = Formative use, S = Summative use.


Table 3.2 General characteristics of the 27 COSs in this review (continued)

<table>
<thead>
<tr>
<th>COS</th>
<th>Full name</th>
<th>Main reference</th>
<th>Subject*</th>
<th>Grade</th>
<th>Teacher/ student</th>
<th>Items</th>
<th>Scale points</th>
<th>Purposeb</th>
<th>Used by others</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTmeter</td>
<td>Competency Thermometer [Competentie Thermometer]</td>
<td>Competentie Thermometer, n.d.-a</td>
<td>✓</td>
<td>K – 12</td>
<td>Teacher</td>
<td>46</td>
<td>3</td>
<td>✓</td>
<td>Yes</td>
</tr>
<tr>
<td>DCO</td>
<td>Diagnostic Classroom Observation</td>
<td>Saginor, 2008</td>
<td>✓ ✓</td>
<td>K – 12</td>
<td>Both</td>
<td>28 / 17</td>
<td>5</td>
<td>✓</td>
<td>Yes</td>
</tr>
<tr>
<td>DIT</td>
<td>Dialogic Inquiry Tool</td>
<td>Reznitskaya, Oyler, &amp; Glina, 2015</td>
<td>✓</td>
<td>4 – 6</td>
<td>Both</td>
<td>19</td>
<td>6</td>
<td>✓</td>
<td>Yes</td>
</tr>
<tr>
<td>DLLT</td>
<td>Developing Language and Literacy Teaching rubrics</td>
<td>Hough et al., 2013</td>
<td>✓</td>
<td>K – 2</td>
<td>Teacher</td>
<td>51</td>
<td>4</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>ECCOM</td>
<td>Early Childhood Classroom Observation Measure</td>
<td>Stipek &amp; Byler, 2004</td>
<td>✓ ✓ ✓</td>
<td>Pre-k – 1</td>
<td>Both</td>
<td>17</td>
<td>5</td>
<td>✓</td>
<td>Yes</td>
</tr>
<tr>
<td>FfT</td>
<td>Framework for Teaching</td>
<td>Danielson, 2013</td>
<td>✓</td>
<td>K – 12</td>
<td>Both</td>
<td>16</td>
<td>4</td>
<td>✓</td>
<td>Yes</td>
</tr>
<tr>
<td>ICALT</td>
<td>International Comparative Analysis of Learning and Teaching</td>
<td>van de Grift, 2007</td>
<td>✓</td>
<td>K – 12</td>
<td>Both</td>
<td>35</td>
<td>4</td>
<td>✓</td>
<td>Yes</td>
</tr>
<tr>
<td>IGDI-model</td>
<td>The extent of working with the Interactive Differentiated Direct Instruction Model [Mate van werken met het Interactief Gedifferentieerd Directe Instructionmodel]</td>
<td>Houtveen, Mijs, Vernooij, Van de Grift, &amp; Koekebacker, 2003</td>
<td>✓</td>
<td>1 – 2</td>
<td>Teacher</td>
<td>28</td>
<td>5</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>IQA</td>
<td>Instructional Quality Assessment</td>
<td>Boston, 2012a</td>
<td>✓ ✓</td>
<td>K – 12</td>
<td>Both</td>
<td>16 / 8</td>
<td>5 / 6</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>COS</th>
<th>Full name</th>
<th>Main reference</th>
<th>Subject</th>
<th>Grade</th>
<th>Teacher / student</th>
<th>Items</th>
<th>Scale points</th>
<th>Purpose</th>
<th>Used by others</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marzano Protocol</td>
<td>Marzano Center Teacher Observation Protocol</td>
<td>Learning Sciences Marzano Center, 2016</td>
<td>✓</td>
<td>K – 12</td>
<td>Teacher</td>
<td>41</td>
<td>5</td>
<td>✓ ✓</td>
<td>Yes</td>
</tr>
<tr>
<td>MECOR</td>
<td>Mathematics Enhancing Classroom Observation Record</td>
<td>Schaffer, Muijs, Kitson, &amp; Reynolds, n.d.</td>
<td>✓</td>
<td>K – 6</td>
<td>Teacher</td>
<td>57</td>
<td>5</td>
<td>✓</td>
<td>Yes</td>
</tr>
<tr>
<td>MQI</td>
<td>Mathematical Quality of Instruction</td>
<td>Hill, 2014</td>
<td>✓</td>
<td>K – 9</td>
<td>Both</td>
<td>31</td>
<td>4 / 5</td>
<td>✓ ✓ ✓</td>
<td>Yes</td>
</tr>
<tr>
<td>QCI</td>
<td>Quality of Classroom Instruction</td>
<td>Doabler, Nelson, et al., 2014</td>
<td>✓</td>
<td>K</td>
<td>Teacher</td>
<td>9</td>
<td>3</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>RATE</td>
<td>Rapid Assessment of Teacher Effectiveness</td>
<td>Gargani &amp; Strong, 2014</td>
<td>✓</td>
<td>3 – 12</td>
<td>Teacher</td>
<td>6</td>
<td>3</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>RCMIS</td>
<td>Ratings of Classroom Management and Instructional Support</td>
<td>Doabler, Nelson, et al., 2014</td>
<td>✓</td>
<td>K</td>
<td>Both</td>
<td>11</td>
<td>4</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>SAMPI</td>
<td>Science and Mathematics Program Improvement Lesson Observation Protocol</td>
<td>SAMPI, 2003a</td>
<td>✓ ✓ ✓</td>
<td>K – 12</td>
<td>Both</td>
<td>38 - 56</td>
<td>3 / 7</td>
<td>✓</td>
<td>Yes</td>
</tr>
<tr>
<td>Science Discourse Instrument</td>
<td>Science Discourse Instrument</td>
<td>Osborne et al., 2014</td>
<td>✓ ✓ ✓</td>
<td>3 – 5</td>
<td>Both</td>
<td>13</td>
<td>5</td>
<td>✓ ✓ ✓</td>
<td></td>
</tr>
<tr>
<td>TAP observational rubric</td>
<td>Teacher Advancement Program observational rubric</td>
<td>TAP, n.d.-a</td>
<td>✓ ✓ ✓</td>
<td>K – 12</td>
<td>Both</td>
<td>19</td>
<td>5</td>
<td>✓ ✓ ✓</td>
<td>Yes</td>
</tr>
<tr>
<td>VTBI</td>
<td>Virgilio Teacher Behavior Instrument</td>
<td>Teddlie, Virgilio, &amp; Oescher, 1990</td>
<td>✓ ✓ ✓</td>
<td>K – 12</td>
<td>Teacher</td>
<td>35</td>
<td>6</td>
<td>✓ ✓ ✓</td>
<td>Yes</td>
</tr>
</tbody>
</table>

The quality of classroom observation systems

in the lesson or not (SAMPI). The items in the COSs are about both teacher and student behavior (16 COSs), or only about teacher behavior (11 COSs). In the DIT, most items are rated twice, once for teacher behavior and once for student behavior, to assess teacher behavior in context and take into account its effect on students.

The reviewers agreed that the purpose of the COS was clearly described by its developers for 25 COSs. Most COSs are meant to be used in research (18 COSs) and/or for formative evaluation aimed at teacher development (19 COSs). For eight COSs, a summative purpose (the assessment of teachers) is also mentioned.

» COS content

All reviewed COSs measure (an aspect of) teaching quality, since doing so was one of the inclusion criteria for this review. Nevertheless, the systems measure different teaching quality constructs or define and/or operationalize teaching quality differently. For 25 COSs, the reviewers agreed that instrument developers provided a clear description of the constructs that the COSs aim to measure.

Table 3.3 provides an overview of which of the nine dimensions of teaching, fundamental to students’ learning (Bell et al., 2018), are included in each of the 27 COSs. A checkmark implies that at least one criterion in the COS measures a specific dimension. The number of teaching dimensions covered by the COSs ranges from one dimension (DIT) to all nine dimensions (TAP), with an average of five dimensions. The dimensions involvement/motivation of students, cognitive activation, safe and stimulating classroom climate, and assessment for learning are measured by most COSs (19-23 COSs), while the quality of subject matter representation, teaching learning strategies and student self-regulation, and instructional differentiation are measured less frequently (9-11 COSs).

» Theoretical basis of the COSs

The theoretical basis for the items in the 27 COSs can be found in (reviews of) the teacher effectiveness literature, (reviews of) literature on instruction, teaching, learning or development, and/or national teaching and learning standards. Sometimes, other input has been used for COS development, such as existing COSs, the analysis of videotapes (MQI), or experiments. RATE was developed based on experiments on how well raters could identify highly effective and ineffective teachers (Strong, Gargani, & Hacifazlioglu, 2011).
### Table 3.3 Nine teaching dimensions in the COSs

<table>
<thead>
<tr>
<th>COS</th>
<th>Climate</th>
<th>Management</th>
<th>Motivation</th>
<th>Explanation</th>
<th>Quality of representation</th>
<th>Cognitive activation</th>
<th>AFL</th>
<th>Differentiation</th>
<th>Learning strategies</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADAPT</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>4</td>
</tr>
<tr>
<td>AIMS Instr.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>5</td>
</tr>
<tr>
<td>ASOS</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>CLASS K-3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>5</td>
</tr>
<tr>
<td>CLASS UE</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>7</td>
</tr>
<tr>
<td>CMI Protocol</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>8</td>
</tr>
<tr>
<td>CSS-Form</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>8</td>
</tr>
<tr>
<td>CTI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>4</td>
</tr>
<tr>
<td>CTmeter</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>5</td>
</tr>
<tr>
<td>DCO</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>7</td>
</tr>
<tr>
<td>DIT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>DLLT</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>7</td>
</tr>
<tr>
<td>ECCOM</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>7</td>
</tr>
<tr>
<td>FfT</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>7</td>
</tr>
<tr>
<td>ICA L T</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>8</td>
</tr>
<tr>
<td>IGDl-Model</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>6</td>
</tr>
<tr>
<td>IQA</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>2</td>
</tr>
</tbody>
</table>

Note:  
- Climate = Safe and stimulating learning climate.  
- Management = Classroom management.  
- Motivation = Involvement/motivation of students.  
- Explanation = Explanation of subject matter.  
- Quality of representation = Quality of subject matter representation.  
- AFL = Assessment for Learning.  
- Differentiation = Differentiated instruction.  
- Learning strategies = Teaching learning strategies and student self-regulation.
Table 3.3 Nine teaching dimensions in the COSs (continued)

<table>
<thead>
<tr>
<th>COS</th>
<th>Climate</th>
<th>Management</th>
<th>Motivation</th>
<th>Explanation</th>
<th>Quality of representation</th>
<th>Cognitive activation</th>
<th>AFL</th>
<th>Differentiation</th>
<th>Learning strategies</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marzano Protocol</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>MECORS</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>MQI</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>QCI</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>RATE</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>RCMIS</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>SAMPI protocol</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Science Discourse Instrument</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>TAP rubric</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>9</td>
</tr>
<tr>
<td>VTBI</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>19</td>
<td>18</td>
<td>23</td>
<td>17</td>
<td>9</td>
<td>22</td>
<td>19</td>
<td>10</td>
<td>11</td>
<td>139</td>
</tr>
</tbody>
</table>

Note: * Climate = Safe and stimulating learning climate. b Management = Classroom management. c Motivation = Involvement/motivation of students. d Explanation = Explanation of subject matter. e Quality of representation = Quality of subject matter representation. f AFL = Assessment for Learning. g Differentiation = Differentiated instruction. h Learning strategies = Teaching learning strategies and student self-regulation.
The reviewers agreed that the basis for the items used in the COS was clear and sufficiently elaborated for 19 COSs. However, the amount of information provided differs considerably between COSs. Some COS developers provide an extensive summary of the literature underlying each criterion in the COS, while others only refer to the literature used in a general sense. For two COSs, the instrument developers only mentioned that the COS was based on extensive literature reviews, with no further specification of the literature (which was evaluated as insufficient by both reviewers).

If the underlying theory is not elaborated enough by the COS developers, it will be difficult for others to judge whether the items in the COS match the theory and/or whether the items cover the constructs well. For 12 COSs, both reviewers judged that the items in the COS match the theoretical basis of the COSs. For the other COSs, this was not clear to at least one of the reviewers.

» Quality of the items
For eight COSs, both reviewers agreed that the items in the COS were problematic (could add to measurement error) and for 11 COSs one of the two reviewers thought this was the case. The most common problem indicated by reviewers was that multiple behaviors were measured with one item.

Rating quality measures
This section shows to what extent instrument developers have taken measures to assure high rating quality and reduce rater error. The section specifically focuses on the provision of a rater manual, the definition of scoring rules, the provision of rater training, the specification of the required rater expertise, the specification of the required number of raters, and the provision of rater instructions.

» Providing a rater manual
Table 3.4 shows whether instrument developers developed a rater manual. Three types of rater manuals could be distinguished. For seven COSs, a rater manual aimed at external users was found. For four COSs, a rater manual was found that was not aimed at external users but that was used in an instrument developers’ research project. For another five COSs, information on how to use the COS was found in other documents, for example, in books or online. For the other 11 COSs, no such information for raters was found. The manual of the VTBI was no longer available.
The quality of classroom observation systems

According to the reviewers (based on their explanations), most manuals include information about the constructs that can be measured with the COS, the teachers (grade) that can be observed with the COS, and the intended use(s) of the COS (formative/summative). The reviewers found that information concerning the required number of raters for a specific use of the COS is provided in only three manuals. A summary of the research findings about the COS (for example, regarding the reliability of the scale or its predictive validity) made by the instrument developers, aimed at external users, was found in five manuals. According to one of the reviewers, this information was rather brief or outdated in two manuals. Finally, (one of) the reviewers found information about the required expertise of raters in seven manuals. In most cases, this entails training and rater certification requirements.

» Providing scorings rules
Scoring rules were found for 21 of the COSs. For all of these COSs, except for the ICALT instrument, this means that for each or most score points on the rating scale, instrument developers have provided a description of the behavior that exemplifies that score. The ICALT does provide examples of quality teaching for all items but not for each score point.

» Providing rater training
Table 3.4 shows that all but one instrument developers mention the training of raters. For half of the COSs, some information on the content of the training is available. Most training sessions include a theoretical component, in which the instrument and its theoretical basis are explained, and a practical component in which observation and rating with the COS are practiced and ratings are discussed. In some training activities, attention is paid to conducting conversations with teachers before and after the observations, so-called pre- and post-conferences. Other training activities include learning to use the gathered information to improve teaching and learning. The total duration of the training activities ranges from two hours (RATE) to nine days (DCO). However, some training activities do not solely focus on the COS but on other instruments in the system or research project as well.

Often, it is unclear if external users can be trained in the use of the COS. For only a few instruments, this is clearly indicated on a website. For example, Teachstone offers a 2-day CLASS training course and MQI offers an online training.
### Table 3.4 Rating quality procedures

<table>
<thead>
<tr>
<th>COS</th>
<th>Rater manual</th>
<th>Scoring rules</th>
<th>Training</th>
<th>Rater expertise</th>
<th>Recertification</th>
<th>Number of raters</th>
<th>Instructions on how to act</th>
<th>Instructions on when to score</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADAPT</td>
<td>External user manual</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>≥ 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIMS instrument</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASOS</td>
<td>Internal research manual</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>1-3*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLASS K-3</td>
<td>External user manual</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>Every year</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLASS UE</td>
<td>External user manual</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>Every year</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CMI Protocol</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSS-Form</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CTI</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>1-2*</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>CTmeter</td>
<td>Other</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DCO</td>
<td>Other</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>DIT</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>3*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DLLT</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>1*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ECCOM</td>
<td>External user manual</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>FFt</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICALT</td>
<td>Internal research manual</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>IGDI-model</td>
<td>External user manual</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: * not indicated by the instrument developers but derived from research by the instrument developers.
Table 3.4 Rating quality procedures (continued)

<table>
<thead>
<tr>
<th>COS</th>
<th>Rater manual</th>
<th>Scoring rules</th>
<th>Training</th>
<th>Rater expertise</th>
<th>Recertification</th>
<th>Number of raters</th>
<th>Instructions on how to act</th>
<th>Instructions on when to score</th>
</tr>
</thead>
<tbody>
<tr>
<td>IQA</td>
<td>External user manual</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>2*</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Marzano Protocol</td>
<td>Other</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>2-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MECORS</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>MQI</td>
<td>Other</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>2</td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td>QCI</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>RATE</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>2*</td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td>RCMIS</td>
<td>Internal research manual</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>SAMPI observation protocol</td>
<td>External user manual</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Science Discourse Instrument</td>
<td>Other</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>n/a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TAP observational rubric</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Every year</td>
<td>&gt; 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VTBI</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>&gt; 1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: *not indicated by the instrument developers but derived from research by the instrument developers.
Chapter 3

» Required rater expertise
A description of the required expertise for raters using the COS was found for 21 COSs (see Table 3.4). For most of these COSs, instrument developers mention that there is an inter-rater reliability threshold that raters should meet after training. This exact threshold is rarely specified. Sometimes, other rater competencies are also measured. For example, the CSS includes a content knowledge assessment for raters, and the TAP requires that raters write a post-conference plan. For three COSs, information on the recommended/required recertification of raters (after one year) was found (see Table 3.4).

» Required number of raters
For six COSs, instructions were found regarding the number of raters that should observe a single teacher (see Table 3.4) when using the COS for a specific purpose. Most often the instructions imply recommendations for multiple raters. No guidelines on this were found for the other 21 COSs. For six of them, the number of raters was specified in the instrument developers’ research (which was also most of the time by multiple raters). For 14 of these 20 COSs, the instrument developers were not explicit about the number of raters in their research papers, which gives the impression that teachers were observed by a single rater. For one COS, no research carried out by the instrument developers was available.

» Rater instructions
Instructions for raters on how to act during the observation were found for six COSs (see Table 3.4). The developers of four COSs (CTI, ECCOM, ICALT, SAMPI) have included in their instructions that raters should fade into the background as much as possible or stay in the periphery of the action, whilst discouraging interaction and asking questions. According to the author of DCO, it is not effective to sit inconspicuously at the back, raters have to take a closer look at what happens in the classroom. The author of the IQA encourages observes to circulate through the classroom during observation. For three COSs, such rater instructions are not relevant because observations are only conducted using video.

Instructions for raters on when to score the items in the COS were found for 17 COSs. Most COSs need to be scored at the end of or after the observation period (13 COSs). For three COSs, raters are instructed to (also) score during the observation, and for one COS the rating is supposed to be done after the post-conference.
Sampling specifications

Table 3.5 shows four observation sampling specifications that the instrument developers can recommend to the external users of a COS. For eight COSs, the instrument developers have provided information on the number of lesson observations to conduct per teacher. This varies from one to six lesson observations. If instrument developers did not provide this information, the table indicates how many observations per teachers the instrument developers conducted in their own research (indicated with an asterisk). This varies from one to eight observations. Six COSs lack both guidelines regarding the number of observations per teacher and research conducted by instrument developers from which the desired number of observations per teacher can be derived.

Most instrument developers do not provide guidelines regarding the recommended observation moment during the day, week, or school year. In three COSs, guidelines indicate that observations can best be conducted at the beginning of the school day. For one COS, the same information could be derived from the research by the COS developers. Vitiello, Hadden, and Teachstone Policy Group (2014) argue that there are some indications that CLASS scores are slightly lower at the beginning of the school year, surrounding the winter holidays, and at the end of the school year, and they recommend to avoid observing then. The developers of RATE describe that RATE is preferably used at the beginning of the school year, to enable teacher trainers, mentors, or school administrators to, based on the RATE data, make appropriate adjustments in teacher support early enough in the school year, thus increasing the likelihood of enhanced student performance on standardized tests.

For seven COSs, instrument developers specify the lesson types that should be observed. This includes a specification of the lesson activities that need to take place in the classrooms (e.g., instruction, working on assignments, discourse) or a specification of the learning activities taking place during the lesson (interacting with/developing new knowledge, or word study).

Most instrument developers specify the required length of the observation, or alternatively, this information could be derived from the research by the instrument developers (indicated with an asterisk in Table 3.5). The recommended observation length varies from 20 minutes to 3 hours, but most observations are supposed to last a single lesson.
### Table 3.5 Sampling specifications

<table>
<thead>
<tr>
<th>COS</th>
<th>Number of observations</th>
<th>Observation moment</th>
<th>Lesson type</th>
<th>Observation length</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADAPT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIMS instrument</td>
<td>2-4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASOS</td>
<td>1-2*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLASS K-3</td>
<td>1</td>
<td>Start of the day, some periods during the year are indicated as less appropriate</td>
<td>Develop, practice, or solidify lessons</td>
<td>1 hour</td>
</tr>
<tr>
<td>CLASS UE</td>
<td>≥ 4</td>
<td>Start of the day, some periods during the year are indicated as less appropriate</td>
<td>At least 3 or 4 cycles of 25 minutes of observation and scoring</td>
<td>At least 4 cycles of 30 minutes of observation and scoring</td>
</tr>
<tr>
<td>CMI Protocol</td>
<td>1*</td>
<td></td>
<td>Develop, practice, or solidify lessons</td>
<td>Whole lesson*</td>
</tr>
<tr>
<td>CSS-Form</td>
<td>≥ 2</td>
<td></td>
<td>Classroom talk, language arts instruction, and working on assignments</td>
<td>30 minutes</td>
</tr>
<tr>
<td>CTI</td>
<td>8*</td>
<td></td>
<td>Classroom discussions of assigned readings</td>
<td>45 minutes*</td>
</tr>
<tr>
<td>CTmeter</td>
<td>3</td>
<td>Start of the day</td>
<td>Includes separate scales for interactive read aloud, shared reading, guided reading, interactive writing, writing workshop, and word study</td>
<td>1.5 hours</td>
</tr>
<tr>
<td>DCO</td>
<td></td>
<td></td>
<td></td>
<td>Whole lesson</td>
</tr>
<tr>
<td>DIT</td>
<td></td>
<td></td>
<td>Classroom discussions of assigned readings</td>
<td>20 minutes*</td>
</tr>
<tr>
<td>DLLT</td>
<td>1*</td>
<td></td>
<td></td>
<td>Full literacy block*</td>
</tr>
</tbody>
</table>

Note: *not indicated by the instrument developers but derived from research by the instrument developers.
### Table 3.5 Sampling specifications (continued)

<table>
<thead>
<tr>
<th>COS</th>
<th>Number of observations</th>
<th>Observation moment</th>
<th>Lesson type</th>
<th>Observation length</th>
</tr>
</thead>
<tbody>
<tr>
<td>ECCOM</td>
<td>1*</td>
<td>Start of the day*</td>
<td></td>
<td>3 hours*</td>
</tr>
<tr>
<td>Fit T</td>
<td></td>
<td></td>
<td></td>
<td>Whole lesson</td>
</tr>
<tr>
<td>ICALT</td>
<td>1*</td>
<td></td>
<td></td>
<td>Whole lesson</td>
</tr>
<tr>
<td>IGDI-model</td>
<td>1*</td>
<td></td>
<td></td>
<td>Whole lesson*</td>
</tr>
<tr>
<td>IQA</td>
<td>1-2*</td>
<td></td>
<td></td>
<td>Whole lesson*</td>
</tr>
<tr>
<td>Marzano Protocol</td>
<td>3</td>
<td></td>
<td>Lessons in which students interact with new knowledge, practice and deepen their understanding of new knowledge, or apply knowledge</td>
<td>Whole lesson</td>
</tr>
<tr>
<td>MECORS</td>
<td>1*</td>
<td></td>
<td></td>
<td>Whole lesson*</td>
</tr>
<tr>
<td>MQI</td>
<td>3-4</td>
<td></td>
<td></td>
<td>6-8 segments of 5 / 7.5 minutes</td>
</tr>
<tr>
<td>QCI</td>
<td>3*</td>
<td></td>
<td></td>
<td>Whole lesson*</td>
</tr>
<tr>
<td>RATE</td>
<td>1*</td>
<td>Beginning of the school year</td>
<td>Direct instruction of new material</td>
<td>20 minutes*</td>
</tr>
<tr>
<td>RCMIS</td>
<td>3*</td>
<td></td>
<td></td>
<td>Whole lesson*</td>
</tr>
<tr>
<td>SAMPI observation prot.</td>
<td></td>
<td></td>
<td></td>
<td>Whole lesson</td>
</tr>
<tr>
<td>Science Discourse Instrument</td>
<td>3*</td>
<td></td>
<td>Lesson with group science discourse</td>
<td>2 x 15 minutes*</td>
</tr>
<tr>
<td>TAP observational rubric</td>
<td>3-6</td>
<td></td>
<td></td>
<td>Whole lesson</td>
</tr>
<tr>
<td>VTBI</td>
<td></td>
<td></td>
<td></td>
<td>Whole lesson*</td>
</tr>
</tbody>
</table>

Note: *not indicated by the instrument developers but derived from research by the instrument developers.
The reviewers evaluated whether the sampling specifications have been described and substantiated sufficiently by the instrument developers, which is described later.

### 3.4.2 Reliability

In the next section, the available information on the reliability of the COSs scores is presented.

**Available reliability information**

For all but one COSs, information about rater reliability was available. A wide range of measures are reported. Table 3.6 shows which measures were found and how frequently the measures were reported in at least one publication concerning the COSs. A generalizability study was found for five COSs.

<table>
<thead>
<tr>
<th>Reported measure</th>
<th>Number of COSs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of exact agreement</td>
<td>15</td>
</tr>
<tr>
<td>Percentage of adjacent (within one point of one another) agreement</td>
<td>10</td>
</tr>
<tr>
<td>Intraclass Correlation Coefficient</td>
<td>9</td>
</tr>
<tr>
<td>Correlation coefficient</td>
<td>8</td>
</tr>
<tr>
<td>Cohen's Kappa coefficient</td>
<td>7</td>
</tr>
<tr>
<td>Weighted Kappa</td>
<td>5</td>
</tr>
<tr>
<td>Mean squared difference</td>
<td>3</td>
</tr>
<tr>
<td>Cronbach's alpha</td>
<td>1</td>
</tr>
<tr>
<td>T-test results</td>
<td>1</td>
</tr>
<tr>
<td>Percentage of scores within .75 point of an expert rater</td>
<td>1</td>
</tr>
</tbody>
</table>

These different measures for rater reliability are hard to compare because some imply absolute rater agreement while others imply rater consistency. Some measures take into account the probability that raters agree based on chance, while others do not. In some publications, multiple measures are reported based on the same rater data. These publications show that whether rater reliability should be judged as sufficient also depends on the measure reported. For example, Junker et al. (2006) reported that the percentage of exact agreement and Cohen’s Kappa indicate insufficient agreement between raters on the IQA (respectively 51% and .33), while the percentage of adjacent
agreement and the intraclass correlation coefficient indicate sufficient agreement (respectively 96% and .74). Park, Chen, and Holtzman (2014) reported multiple measures for inter-rater reliability for CLASS UE, FfT, and MQI. The percentage of adjacent agreement is over 75% for all three COSs, while the Cohen’s Kappa values are .21, .24, and .51 respectively. Often, authors report only a single measure, which is often a measure that does not take into account the probability of agreeing by chance. For 11 COSs, no publications were found that report measures that take into account such chance.

For most COSs information about the reliability of the COS scale was found. For 20 COSs, a Cronbach’s alpha was reported in at least one publication. Test-retest reliability was reported in the publications involving two COSs. For one COS, we found a split-half correlation and a Spearman-Brown correlation. For two COSs, the results of a Rash analysis were found.

**Evaluation of reliability**

For only five COSs, both reviewers indicated that the available information allowed a thorough judgment of the reliability of the COS scale. For these instruments, information about inter-rater reliability and the reliability of the scale was available. For all of these five COSs, the reviewers evaluated the findings of the reliability research as sufficient with respect to the intended type of decisions (either formative and/or summative) made based on the COSs. The quality of the reliability research was also evaluated positively by both reviewers for these five COSs. The procedures followed seemed to be correct and the samples of teachers in the reliability research were representative of the teachers for whom the COSs had been designed for (one reviewer disagreed for one COS on this aspect).

For 16 COSs, the reviewers agreed that a thorough reliability judgement was not possible using the available information. Although there was some information available about the inter-rater reliability and/or the reliability of the scale for these COSs, more reliability research is needed. The reviewers also agreed that the findings of the reliability research were insufficient to support the use of COS scores for the intended types of decisions for eight of these COSs, that the procedures for computing the reliability coefficients were incorrect for seven COSs, and that the sample for computing the coefficients were not consistent with the intended use of the COSs for four of these COSs.
For the remaining six COSs, the reviewers disagreed about whether there was enough information to make a thorough reliability judgement.

3.4.3 Validity argument

The reviewers evaluated the evidence for the valid use of the COSs by reviewing the scoring inference and, if applicable, also the generalization, extrapolation, and the implication inference in the validity argument. Whether the use of a COS will result in valid scores cannot be evaluated in a review such as the one carried out here, as this depends on the user context, how the COS is implemented, and how the scores are used. What reviewers did do in this review is evaluate the characteristics of the COSs that contribute to its valid use, as well as the research supporting the valid use of the COSs. For this purpose, the reviewers evaluated the same warrants and backing for each COS, while taking into account the intended use and characteristics of each specific COS as specified by the instrument developers. Although the inferences are related, the evaluation of each inference was conducted separately. Generalizing a score cannot be valid if the scoring inference is presumed to be invalid. However, the warrants for the generalization and extrapolation inferences were formulated in such a way that they were relevant to evaluate, even in the absence of evidence for other inferences.

Scoring inference

In order to evaluate the scoring inference, the reviewers evaluated three warrants.

» The scoring rule(s) is/are substantiated (statistically)

The first warrant concerns the scoring rule(s), which was/were defined in the review as scoring rules at the item level (indicating when to assign a specific score) and scoring rules for computing a composite score (for example, the mean of all scale scores). Table 3.7 shows the results of the evaluation of six backings for this warrant. The first four backings were scored if scoring rules were available (26 COSs). For three COSs, one of the reviewers indicated that it lacked scoring rules, therefore the data for these three COSs are based on the evaluation by one reviewer. The last two backings were only scored if rules for computing a composite score were available (23 COSs).

The results show that, according to at least one of the reviewers, most COSs are based on theory, research, and/or national standards. For half of the COSs, at least one reviewer indicated that the scoring rule(s) is/are supported by experts. For most of
The quality of classroom observation systems

Table 3.7 Evaluation of the scoring inference

<table>
<thead>
<tr>
<th>Warrant 1: The scoring rule(s) is/are (statistically) substantiated (n=26)</th>
<th>Negative</th>
<th>No consensus</th>
<th>Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Freq.</td>
<td>%</td>
<td>Freq.</td>
</tr>
<tr>
<td>The COS is based on theory, research and/or standards (n=26)</td>
<td>8</td>
<td>31%</td>
<td>11</td>
</tr>
<tr>
<td>The scoring rule(s) is/are supported by experts (n=26)</td>
<td>13</td>
<td>50%</td>
<td>4</td>
</tr>
<tr>
<td>The scoring rule(s) is/are supported by teachers (n=26)</td>
<td>21</td>
<td>81%</td>
<td>1</td>
</tr>
<tr>
<td>The scoring rule(s) has/have been tested in (pilot) research (n=26)</td>
<td>8</td>
<td>31%</td>
<td>7</td>
</tr>
<tr>
<td>Statistical analyses support the scoring rule(s) (n=23)</td>
<td>8</td>
<td>35%</td>
<td>3</td>
</tr>
<tr>
<td>The psychometric quality of the items is sufficient (n=23)</td>
<td>13</td>
<td>57%</td>
<td>6</td>
</tr>
<tr>
<td>Warrant 2: Measures were taken to score accurately and consistently (n=27)</td>
<td>6</td>
<td>22%</td>
<td>10</td>
</tr>
<tr>
<td>For each criterion, scoring rules are available for raters (n=27)</td>
<td>6</td>
<td>22%</td>
<td>7</td>
</tr>
<tr>
<td>Raters are trained in the use of the COS (n=27)</td>
<td>1</td>
<td>4%</td>
<td>6</td>
</tr>
<tr>
<td>Raters are expected to meet a certain level of expertise (n=27)</td>
<td>6</td>
<td>22%</td>
<td>11</td>
</tr>
<tr>
<td>Inter-rater reliability is sufficient (n=27)</td>
<td>6</td>
<td>22%</td>
<td>8</td>
</tr>
<tr>
<td>Observation scores are consistent over time (n=27)</td>
<td>19</td>
<td>70%</td>
<td>4</td>
</tr>
<tr>
<td>Warrant 3: Attention is dedicated to preventing rater bias (n=27)</td>
<td>20</td>
<td>74%</td>
<td>5</td>
</tr>
<tr>
<td>Attention is paid to preventing rater bias during rater training and/or in the COS manual (n=27)</td>
<td>24</td>
<td>89%</td>
<td>3</td>
</tr>
<tr>
<td>Multiple raters are used to compute an observation score (n=27)</td>
<td>16</td>
<td>59%</td>
<td>6</td>
</tr>
<tr>
<td>Statistical analyses show that raters do not rate specific groups of teachers differently from others (n=27)</td>
<td>25</td>
<td>93%</td>
<td>1</td>
</tr>
<tr>
<td>Inter-rater reliability is sufficient (n=27)</td>
<td>6</td>
<td>22%</td>
<td>8</td>
</tr>
<tr>
<td>Scoring inference (n=27)</td>
<td>13</td>
<td>48%</td>
<td>9</td>
</tr>
</tbody>
</table>

Note. Due to rounding of the percentages, the percentages do not always add up to 100%
these COSs, experts were consulted during the development of the system in focus groups or using a questionnaire. Often (21 of the 26 COSs), there was no indication of the consultation of teachers (the target group of the COSs). A few COS developers consulted teachers (and school leaders) during the development or use of the system to investigate how teachers experienced the system. For most COSs (18 of the 26 COSs), at least one reviewer indicated that (pilot) research concerning the scoring rules was available. Reviewers disagreed relatively often on this point because it was unclear to them whether the research provided information on the scoring rules. When COSs include a scoring rule to compute a composite score (23 COSs), statistical support for this rule was found by at least one of the reviewers for 15 COSs. Support for the psychometric quality of the items was found less often, for 10 of the 23 COSs support was found by one of the reviewers.

Based on this backing, the reviewers agreed that for 7 of the 26 COSs, the scoring rule(s) was/were substantiated (statistically). For eight COSs, the reviewers, based on the available information, were not convinced of this. The reviewers did not agree on the other 11 COSs.

» Measures were taken to score accurately and consistently

The second warrant concerns the measures that were taken to score accurately and consistently. Table 3.7 shows the results of the evaluation of the six backings for this warrant.

Table 3.4 shows that scoring rules on the item level were available for 21 COSs. However, for seven of these COSs, one of the reviewers rated the scoring rules negatively, for example, because the scoring rules were deemed to provide insufficient information. As can be seen in Table 3.4, raters are trained for all but one COS. However, for six COSs, one of the reviewers evaluated the training activities as an insufficient backing of this warrant. There were various reasons for this, such as a lack of information about the training, the fact that the training is not compulsory, and the belief that the training is too short to qualify as backing for this warrant. Table 3.4 also shows that for 21 COSs at least one reviewer indicated that some level of expertise is expected from raters in the form of either a minimum percentage of agreement after the training course or that the rater(s) holds a specific position such as school leader. Reviewers disagreed relatively often on the evaluation of this backing, probably because this information was often hard to find or unclear.
For 21 COSs, at least one of the reviewers found research concerning the inter-rater reliability, however, this research was not considered a sufficient backing for the warrant for eight of these COSs by one of the reviewers. One reason for this was the quality of the research. For example, some of the research had been conducted using a very small sample or used procedures that had not been described clearly. For most COSs, no research was available on whether raters score consistently over time.

Based on the evaluation of this backing, the reviewers argued that for 11 COSs sufficient measures have been taken to allow for accurate and consistent scoring. Based on the available information, both reviewers were not convinced of this for six COSs. The reviewers did not agree on the other 10 COSs. From the reviewers’ explanations, it appears that in most of these cases the reviewers were uncertain of their own decision.

» Attention is paid to preventing rater bias

The third warrant concerns the prevention of rater bias. Table 3.7 shows the results of the evaluation of four backings for this warrant.

For most COSs (24 COSs), the reviewers did not find information concerning rater bias in the rater manual or in the content of the rater training course. However, usually information regarding the content of the training was limited. For 11 COSs, at least one of the reviewers thought that teachers were observed by multiple raters. This information was not always clear, because instrument developers did not always specify the number of raters or advised to use multiple raters but did not do so in their own research. The reviewers searched for research into rater bias but only two such studies were found. Reddy, Fabiano, Dudek, and Hsu (2013a) used the partial correlation model for differential item functioning (DIF) of Reynolds and Carson (2005) to identify potential bias in CSS-items. The results showed that items functioned similarly across teachers’ age, degree, and years of teaching experience. Park, Chen, and Holtzman (2014) used logistic regression to examine whether the interactions between characteristics of teachers and raters were associated with measures of agreement for the COSs CLASS UE, FFT and MQI. No significant effects were found. For the evaluation of this warrant, the research into inter-rater agreement was also taken into account.

Based on the evaluation of this backing, the reviewers agreed that there was insufficient evidence that instrument developers paid attention to rater bias for 20 COSs.
Chapter 3

» Overall evaluation of the scoring inference

Based on the evaluation of these three warrants, the reviewers gave an overall evaluation of the scoring inference, taking into account the specifications and the purpose of the COSs. Both reviewers were convinced that scoring, using the procedures specified by the instrument developers, will lead to valid observation scores for only five COSs. This does not mean that scoring with the other COSs cannot be valid, however, based on the evidence that the reviewers evaluated, at least one of the reviewers was not convinced that observing with the COSs will lead to valid observation scores for the purpose described by the instrument developers. For 13 COSs, both reviewers were not convinced, while they disagreed regarding the other nine COSs. In half of these cases, they both were not completely certain of their own evaluation, as could be deduced from the reviewers’ explanations.

Generalization inference

In order to evaluate the generalization inference, the reviewers evaluated one warrant on whether the COSs include clear guidelines on the generalization of the observed score, and thus, whether COS developers are explicit about generalization possibilities. For all COSs, evaluation of this inference seemed relevant, since teachers were observed with the COSs to draw general inferences, not only regarding the observed lessons. In some cases, it was not that clear whether the observed score was generalized, for example, if feedback was provided on the observed lesson and it was unclear whether the COS developers perceived the observed lesson as an indication of lesson quality in general. The generalization inferences was also evaluated for these COSs. For three COSs, one of the reviewers indicated that generalization was not the intention of the instrument developers and therefore the results for those three COSs are based on one reviewer.

» Opportunities for generalizations are explicitly described in the COS

Table 3.8 shows the results of the evaluation of seven backings for this warrant. The first five backings are about whether the COS developers provide sampling specifications for external users regarding the number of observations, the length of the observation, the number of raters, the lesson type, the observation moment during the day, week or year. The specifications provided were already presented in Table 3.4 and Table 3.5. The reviewers evaluated here whether the specifications were provided and substantiated by the COS developers.
Table 3.8 Evaluation of the generalization inference

<table>
<thead>
<tr>
<th>Warrant: Opportunities for generalizations are explicitly described in the COS (n=27)</th>
<th>Negative</th>
<th>No consensus</th>
<th>Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freq.</td>
<td>%</td>
<td>Freq.</td>
<td>%</td>
</tr>
<tr>
<td>18</td>
<td>67%</td>
<td>8</td>
<td>30%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The required number of teacher observations per teacher is specified and substantiated (n=27)</th>
<th>Negative</th>
<th>No consensus</th>
<th>Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freq.</td>
<td>%</td>
<td>Freq.</td>
<td>%</td>
</tr>
<tr>
<td>14</td>
<td>52%</td>
<td>12</td>
<td>44%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The observation length is specified and substantiated (n=27)</th>
<th>Negative</th>
<th>No consensus</th>
<th>Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freq.</td>
<td>%</td>
<td>Freq.</td>
<td>%</td>
</tr>
<tr>
<td>7</td>
<td>26%</td>
<td>13</td>
<td>48%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The number of raters is specified and substantiated (n=27)</th>
<th>Negative</th>
<th>No consensus</th>
<th>Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freq.</td>
<td>%</td>
<td>Freq.</td>
<td>%</td>
</tr>
<tr>
<td>21</td>
<td>78%</td>
<td>5</td>
<td>19%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The observation moment is specified and substantiated (n=27)</th>
<th>Negative</th>
<th>No consensus</th>
<th>Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freq.</td>
<td>%</td>
<td>Freq.</td>
<td>%</td>
</tr>
<tr>
<td>17</td>
<td>63%</td>
<td>8</td>
<td>30%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The lesson type is specified and substantiated (n=27)</th>
<th>Negative</th>
<th>No consensus</th>
<th>Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freq.</td>
<td>%</td>
<td>Freq.</td>
<td>%</td>
</tr>
<tr>
<td>7</td>
<td>26%</td>
<td>11</td>
<td>41%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A generalizability study, a reliability study, or IRT analyses has shown that the sample of observations is representative of the assessment domain (n=27)</th>
<th>Negative</th>
<th>No consensus</th>
<th>Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freq.</td>
<td>%</td>
<td>Freq.</td>
<td>%</td>
</tr>
<tr>
<td>23</td>
<td>85%</td>
<td>2</td>
<td>7%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Research into the variation of observed lessons supports the generalizations of the observed score (n=27)</th>
<th>Negative</th>
<th>No consensus</th>
<th>Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freq.</td>
<td>%</td>
<td>Freq.</td>
<td>%</td>
</tr>
<tr>
<td>18</td>
<td>67%</td>
<td>9</td>
<td>33%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Confidence intervals are available and are based on a convention (n=27)</th>
<th>Negative</th>
<th>No consensus</th>
<th>Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freq.</td>
<td>%</td>
<td>Freq.</td>
<td>%</td>
</tr>
<tr>
<td>26</td>
<td>96%</td>
<td>1</td>
<td>4%</td>
</tr>
</tbody>
</table>

Note. Due to rounding of the percentages, the percentages do not always add up to 100%

Table 3.8 shows that according to at least one of the reviewers, most sampling specifications were not specified and substantiated sufficiently by most instrument developers. Reviewers disagreed relatively often. The reason is probably that it was not very clear to them when a sampling specification was substantiated enough. Instrument developers could describe it, motivate it, or back it up by research, however the latter is scarce. Examples of such research are Hill et al. (2012), concerning the required number of observations and raters for MQI, Joe et al.’s research (2014) into the length of observations for CLASS, FfT, and MQI, and the research by Pianta, La Paro, & Hamre (2008) into the stability of CLASS scores across observation cycles, weekdays, and the school year.

For four COSs, a G-study was found. If the studies were not conducted by the instrument developers, than the results did not always lead to sampling specifications.
for external users. For example, a G-study was conducted in the MET study for CLASS (Bill & Melinda Gates foundation, 2012), but the results did not lead to a substantiation of the sampling specifications in the CLASS UE manual (Pianta et al., 2012). For some COSs, research into the stability of the scores was found. However, this was never seen as convincing evidence of the scoring inference by both reviewers. Confidence intervals for the scores were rarely found.

Based on the backing, the reviewers evaluated the warrant on whether opportunities for generalization are explicitly described in the COS. Only one COS (MQI) was evaluated positively by both reviewers. For eight other COSs, the reviewers disagreed. Based on the explanations by the reviewers, this was caused by the lack of strict guidelines in the evaluation framework for reviewers regarding the extent to which the sampling specifications should be based on research. For the other 18 COSs, the reviewers agreed that generalization opportunities were not described explicitly.

**Extrapolation inference**

The reviewers indicated whether an evaluation by the reviewers of the extrapolation inference was appropriate, since instrument developers did not always intend the extrapolation of the score to a broader domain. Reviewers evaluated the extrapolation inference if instrument developers suggest a broader interpretation of the observed score. For example, if general conclusions about teaching quality are drawn, whereas only part of the construct was measured with the COS. Reviewers also evaluated the extrapolation inference if instrument developers explored the correlation with other measures in a broader domain. For six COSs, both reviewers evaluated the extrapolation inference, for another eight COSs only one of the reviewers did so. Reviewers evaluated one warrant for this issue, namely whether the score in the assessment domain was related to the broader target domain.

» The score in the assessment domain is related to the broader target domain

The three backings that were evaluated for this warrant (see Table 3.9) evaluate both analytical and empirical evidence, as suggested by (Kane, 2006). The reviewers were positive about the relation with the broader target domain for 8 of the 14 COSs. For the other COSs, the reviewers indicated that empirical evidence was lacking or that insufficient information was present to judge this warrant.
### Table 3.9 Evaluation of the extrapolation inference

<table>
<thead>
<tr>
<th>Warrant: The score in the assessment domain is related to the broader target domain (n=14)</th>
<th>Negative</th>
<th>No consensus</th>
<th>Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freq.</td>
<td>%</td>
<td>Freq.</td>
<td>%</td>
</tr>
<tr>
<td>5</td>
<td>36%</td>
<td>1</td>
<td>7%</td>
</tr>
<tr>
<td>The assessment domain covers a great portion of the target domain (face validity) (n=14)</td>
<td>3</td>
<td>21%</td>
<td>11</td>
</tr>
<tr>
<td>The theoretical framework underlying the COS fits within the target domain (n=14)</td>
<td>2</td>
<td>14%</td>
<td>3</td>
</tr>
<tr>
<td>The observed score is related to other measures within the target domain (n=14)</td>
<td>3</td>
<td>21%</td>
<td>2</td>
</tr>
</tbody>
</table>

Note. Due to rounding of the percentages, the percentages do not always add up to 100%

### 3.5 Conclusion and discussion

Many authors point to issues that need to be considered when selecting and/or using a COS (e.g., Bell et al., 2012; Cohen & Goldhaber, 2016; Hill, Charalambous, & Kraft, 2012; Hill, Charalambous, Blazer, et al., 2012; Nava et al., 2018; Sandilos, 2012; van der Lans, van de Grift, van Veen, & Fokkens-Bruinsma, 2016). Generating valid and reliable scores by means of a COS is not self-evident. However, as an overview of the developed COSs and their quality was lacking, users were hampered in making deliberate choices with respect to which COS to use in their own context. From a scientific perspective, it is also valuable to evaluate the extent to which COSs meet the general criteria for evaluation instruments. Therefore, the research questions for this review were: Which COSs have been developed to measure teaching quality in primary education, what is the quality of the COS materials and what evidence is available regarding the reliability and validity of the scores these COSs produce?

In total, 27 of the found COSs met the inclusion criteria for this review. These 27 COSs differ enormously in their proposed use, scoring tools, rating quality procedures, sampling specifications, and the empirical evidence for reliability and validity of the COS-scores. Users should therefore be aware of these differences and the need to clarify for themselves which aspects of teaching quality they find important to evaluate, and which criteria the selected COS should meet.

Information about the features of the COSs, if available, is often not presented in an accessible way by instrument developers, making it difficult for potential COS-users to obtain an overview of the qualities of available COSs and decide which COSs best
fits their own context and intended use.

Our results show that the reviewers were most positive about the quality of the COS materials. The scoring tools in most COSs are based on theory, research, and/or national standards, and in all COSs at least some attention is paid to the rating quality measures. Not all COSs had been developed for use by external users, but for those that had been, often some information on the rating quality measures were lacking. Most COSs also provided few sampling specifications.

For most COSs, at least one of the two reviewers was negative about the availability of empirical evidence regarding the reliability and validity of scores. If such information was available, the reviewers often questioned the quality of the research, for half of the reviewed COSs at least one of the reviewers indicated that the procedures for calculating the reliability coefficients were incorrect, that information regarding the research procedures was lacking, and/or that the findings of the reliability research were insufficient for the intended decision-making. This finding is remarkable as these COSs were included in this review because of the availability of empirical evidence. This could have occurred because no inclusion criteria were set regarding the type of research that had to be available (e.g., research concerning the scale) or the quality of the research.

There were no COSs for which both reviewers indicated that a thorough judgement of its reliability was possible and for which both reviewers were also positive about the evidence of reliability, scoring, generalization and, if applicable, extrapolation. If we only look at the first three aspects, and leave the evidence of generalization and extrapolation out of consideration, then the evidence for reliability and the scoring inference was positively evaluated by both reviewers for two COSs and by one of the two reviewers for another 8 COSs. Most of the COSs in this review that are used internationally fall into this category, however two widely used COSs were reviewed negatively by both reviewers on these aspects in the evaluation framework.

Our findings do not provide a very positive picture regarding the quality of most currently available COSs for measuring teaching quality in primary education. This raises a question about the factors that may have caused this. In our opinion, the following factors might offer an explanation:

- There was no perceived need to make the COS-materials accessible to other users as not all COSs in this review were developed for use by third parties (for example,
The quality of classroom observation systems

some were developed by researchers for use in their own research project) and/or instrument developers were not aware that their COS could be of use to others.

- The required resources, such as time and financing, were insufficient to thoroughly develop and research the COS, to develop materials for external users, to make COS materials accessible, and to keep the information on the COS updated.
- Instrument developers are not aware of the full complexity of COS development and of the research required to provide valid COS scores to teachers.
- Instrument developers are not aware of the information COS-users need for proper COS-use and/or to decide whether the COS is useful in their own context.
- Conventional standards for research into the qualities of COSs are not available.

We will describe three measures that could counteract these factors. First, the creation of an international database with all COSs that have been developed (for a specific level of education and subject), their materials, and the research into them, could make COS developers more aware of the value of their COS for others. Moreover, if COS materials are accessible to others, COSs can also be researched and developed further, even if the instrument developers themselves did not have the resources to do so. The availability of a database of COSs could also help reduce the need for instrument developers and researchers to constantly develop new COSs, leading to the use of relatively uninvestigated instruments (and as a result of which it is, for example, unknown how many lessons should be observed to obtain a reliable score of a teacher’s teaching quality).

Second, an assessment system, with quality criteria and research standards for COSs, designed to evaluate the quality of COSs and the empirical evidence for the reliability and validity of COS-scores could make instrument developers more aware of the complexity of COS development and the research that is needed (for potential users). It could also make it easier for potential COS-users to compare (the empirical evidence of) COSs. Examples of such assessment systems, developed to increase the quality of instruments used in a particular field, are that of The Dutch Committee on Tests and Testing (COTAN, Evers et al., 2010) for the assessment of the quality of psychological tests and that of the Research Center for Examinations and Certification (RCEC; Eggen & Veldkamp, 2012) for the assessment of tests and exams in education. An example of international standards are those of the AERA, APA, and NCME (1999) for educational and psychological testing. The evaluation framework for this review was a first step in developing such an assessment system/standards for COSs, however, based on our experiences throughout this review, we
Chapter 3

think that more attention should be paid to the developing the research standards used in such a framework: what kind of empirical backing for COSs is essential and required? For example, what statistical analysis should be conducted to measure the reliability of raters and what measures should be reported? In order for such an assessment system to succeed, it should be widely supported by COSs developers and researchers. It would also be relevant to include quality criteria that are valued by practitioners who use the systems, such as school leaders and the teachers observed using the COSs.

Third, independent assessments of the COSs using the assessment system can provide potential users with the information they need to make a well-informed choice for a COS. International projects have shown that COSs can function differently in different countries (e.g., ISERP; Teddlie, Creemers, Kyriakides, Muijs, & Yu, 2006). Therefore, national governments could take the role of stimulating the reviewing of COSs by independent partners in their own country, as the Dutch COTAN and RCEC do for (psychological) tests. As a next step, governments might demand that schools only use instruments that comply with the quality criteria within the assessment system for COSs.

3.5.1 Limitations

As the Cohen's Kappa indicates, there was quite some disagreement between the reviewers. Although scoring rules were formulated within the evaluation framework, the review was still dependent on the professional opinion and judgement of the reviewers. This was especially the case for the validity component of the evaluation framework, in which different types of backing can be sufficient evidence of a warrant and in which this review is also dependent on the intended use of the COS. However, reviewers also disagreed on straightforward questions, such as whether a user manual was available because of the number of materials available and reviewers had to decide on what qualified as a manual. The fact that so many different documents were available for each COS complicated the review. Even though the information was structured in the descriptive framework, it was still quite difficult for reviewers to obtain an overview of all the available information. Occasionally, the results in different sources were also contradictory, making an overall judgement difficult. At the same time, there was also often a lack of information, for example, on the research procedures.
A different quality assessment method, such as the COTAN approach (Evers et al., 2010) in which reviewers discuss the criteria that they disagree on to achieve consensus, might have been more appropriate. Sometimes, the explanations of the two reviewers were similar, but their final assessment differed. From other explanations, it was clear that one of the reviewers overlooked some information and as a result judged a COS differently. This could have been avoided using the COTAN method, however we do not expect that the conclusions would have been very different. Although the COTAN method was not feasible in this review, it might be a preferable strategy for similar future reviews.

Another limitation of this review is the timeframe in which the review was conducted. The information was gathered in 2014 and 2015. Screening the information, describing the information in the descriptive framework, reviewing the materials with multiple reviewers took years. Since multiple COSs mentioned in the review were still under development, this review could not include all the information that is now available on the COSs. However, there are no reasons to believe that the conclusions of this review do not apply anymore.

Footnotes
1. The 2014 edition of the Standards for Educational and Psychological Testing was not available yet during the development of the framework. The areas that received particular attention in the 2014 revision are not relevant for our framework. The information relevant for the development of the framework does not differ substantially between both versions.
2. The COTAN criteria are used by the Dutch Committee on Tests and Testing (COTAN) to assess the quality of psychological tests available in the Netherlands. COTAN has audited over 750 tests published for professional use.
Quality lies in the eye of the beholder: a comparison of the perspectives on teaching quality of external raters, students, and teachers
Chapter 4

Abstract

Three frequently used methods for measuring (perceptions of) teaching quality are classroom observations by external raters and student and teacher surveys. Little is known about how these methods relate, whereas such knowledge may have important implications for measuring teaching quality. In this study, the teaching quality scores awarded by external raters, students, and teachers were compared. Data were collected in 25 grade nine mathematics classrooms. Students’ and teachers’ perceptions were measured for three lessons per participating teacher, and these lessons were also rated by external raters. The results showed only a significant but low correlation between external raters and students. Thus, teaching quality scores will differ depending on which of the three methods is used. Those who aim to measure teaching quality should therefore carefully select one or combine multiple methods.

This chapter is a modified version of the manuscript:
4.1 Introduction

Measuring teaching quality is important to a broad array of educational stakeholders (Hill, Umland, Litke, & Kapitula, 2012), because such measures can guide the improvement of teaching and support human resource decisions (Haertel, 2013). This is important, because teachers have been identified by many researchers as the most critical, malleable within-school influence on student learning. Teacher differences account for about 10% of the variance in student test score gains during a single school year (Haertel, 2013; Nye et al., 2004). Measuring teaching quality is also important for teacher certification (Hill et al., 2012), and should be a key component of studying the implementation and effects of interventions that target classroom instruction (Desimone, Smith, & Frisvold, 2010).

There is considerable debate about the best way to measure teaching quality (Chetty et al., 2011). Three common methods for measuring the quality of classroom practice are classroom observation, teacher surveys, and student surveys (Lawrenz, Huffman, & Robey, 2003). Each method has its own strengths and weaknesses. Classroom observation is generally viewed to be the most objective method to measure teaching quality (Lasagabaster & Sierra, 2011). These observations can provide meaningful feedback to teachers if an expert’s opinion is given regarding the quality of classroom practices (Cohen & Goldhaber, 2016; Lawrenz et al., 2003). However, obtaining valid scores through classroom observation is not self-explanatory. Many authors (e.g., Bell et al., 2012; Hill, Charalambous, & Kraft, 2012; Nava et al., 2018) point to issues that need to be taken into account when using classroom observation, for example the number of (trained) external raters that should observe a single teacher and the number of lessons per teacher that should be observed, to obtain a valid score (for an overview of these issues, see chapter 2). This makes classroom observation a time-consuming and costly method for measuring teaching quality.

An easier to administer, cost-effective, and non-intrusive method is the use of student perceptions through questionnaires. Students are the most important stakeholders, spend the most time in the classroom, and know best how teaching is experienced by the ‘client’ (Dockterman, 2017). Seeing classroom observation as the most objective measure reflects a positivist/realist view of social life, while from a symbolic interactionist perspective the beliefs of participants are perceived to be most important (Lawrenz et al., 2003). Since students are the consumers and direct recipients of classroom teaching, they should play a role in its evaluation.
(Dockterman, 2017). Their perceptions can also function as a valuable source of feedback to teachers (Bijlsma et al., n.d.). Disadvantages of student perceptions are that not all students can fill in questionnaires (for example, young children) and perceptions are believed to be influenced by both student variables (e.g., ethnicity; Levy, Brok, Wubbels, & Brekelmans, 2003) and teacher variables unrelated to teaching efficacy (e.g., teacher popularity; Fauth, Decristan, Rieser, Klieme, & Büttner, 2014). However, there are several studies that support the reliability and validity of student ratings (Dockterman, 2017; Marsh, 2007).

Another cost-efficient method for measuring teaching quality is the use of teacher perceptions through questionnaires, that can provide insight into teachers’ intentions, thought processes, knowledge, and beliefs. It also seems important to involve teachers in their own evaluation process, because they have greater knowledge of their classroom context, and can therefore provide insights that an outside rater may not recognize (Goe et al., 2008). Teacher surveys can also stimulate teachers to reflect on their own practices (Dockterman, 2017), however underperforming teachers might overestimate themselves and lack the metacognitive competence to recognize their own underperformance whereas high performing teachers may underestimate their professional skills (Kruger & Dunning, 1999). Self-report may be prone to social desirability, however this may be partly mediated by ensuring the confidentiality of teachers’ responses, gathering longitudinal data, and gathering data from multiple sources (Goe et al., 2008). Another issue with self-report, and also student perceptions, is the importance of ensuring the consistent interpretation of terminology and shared understanding of what is measured. This is less of an issue with classroom observation if external raters are trained in the interpretation of the items.

An interesting question is how these three methods for measuring teaching quality relate to each other. Answering this question will contribute to our knowledge of the three methods and whether they provide the same information. The answer to this question may also have practical implications for the measurement of teaching quality. Most of the time, only one of the three methods is used in research and practice. If congruence between the three methods is low, then the results are dependent on the method chosen: teachers will provide a different picture of teaching quality than external raters or students. Conversely, if the three methods correlate highly, it might be possible to replace classroom observation partly by more cost-efficient methods such as student or teacher surveys.
Little is known about the relation between the three methods within secondary education. Some studies have explored the agreement between the perceptions of teachers and students or between external raters and students, but studies that look into the agreement between all three different methods are, to our knowledge, limited.

In the next section, we provide an overview of recent research that is available about the congruence of the different methods for measuring teaching quality. Only research that was conducted in upper-elementary and secondary education is summarized. Secondary education is the focus of our study, and the context of lower primary school grades is too different from our context. The context in higher education is also different, for example, higher education students are more used to providing their opinions and student-teacher relationships are often more distant in higher education.

4.2 The agreement between three methods for measuring teaching quality

4.2.1 Students’ perceptions compared to teachers’ perceptions
Several studies have investigated the agreement between students’ perceptions and teachers’ perceptions of teacher behavior in the classroom. Quite a number of studies report on differences between student and teacher perceptions of interpersonal teacher behavior in secondary education. Overall, teachers in these studies had a more favorable perception of their own interpersonal behavior than students (den Brok, Bergen, & Brekelmans, 2006). For example, in research by Rickards and Fisher (2000), teachers reported higher ratings than their students regarding their own leadership, helpfulness/friendliness, and understanding (all behaviors positively related to student achievement and motivation; den Brok, Brekelmans, & Wubbels, 2004).

The research in secondary education into the agreement of students’ and teachers’ perceptions of teachers’ instructional behavior provides an ambiguous picture and does not allow general conclusions. In research by den Brok et al., (2006), the degree of divergence differed per scale. On some scales teachers score higher (e.g., on strong and shared control), on others lower (e.g., classroom management). Wagner et al., (2016) found high correlations between students’ and teachers’ perceptions of classroom management, but a moderate correlation for goal clarity, and no correlation for support of autonomy. According to Kunter and Baumert (2006) students did not
seem to be able to differentiate between different instructional approaches.

Research in which teachers and students were asked to report the frequency of classroom activities does not provide a clear picture either. Desimone et al. (2010) report correlations regarding the perceived frequency of mathematics activities in the classroom. Correlations between students and teachers differed from low to high. Similar research by Dockterman (2017) showed that according to students, the instructional practices were rather traditional, while teachers reported more reform-oriented instructional practices, such as small group discussion and experimentation.

Several explanations for the low agreement between students and teachers are mentioned in the literature. Various student and teacher characteristics are mentioned that might influence the perceptions, and thus the correlation between the two. Differences between students’ and teachers’ perceptions of interpersonal classroom behavior may be related to teacher experience (Brekelmans, Wubbels, & den Brok, 2002), ethnicity and the cultural background of teachers and students (den Brok, Levy, Rodriguez, & Wubbels, 2002). Perceptions of students and teachers of instructional behavior might be related to teachers’ instructional style (den Brok et al., 2006), individual student traits (e.g., self-efficacy or self-regulation), or student states (e.g., interest or perceived individual teacher attention; Wagner et al., 2016).

Kunter & Baumert (2006) suggest that the low agreement between students and teachers might be explained by the fact that students’ and teachers’ perceptions tap different aspects of the classroom environment, rather than the same underlying construct. Even if teachers and students answer similarly worded items, their interpretation of the items might differ. In their own research, factor analyses indicated that teacher and student ratings did not share a common conceptual structure.

Wagner et al., (2016) pointed out that, in the vast majority of the studies, the measurement of teachers’ and students’ perceptions occurs only once. A problem with this approach is that the reference period of the respondents might be different during the rating process. Weinstein (as cited in Wagner et al., 2016) noted that when asked about typical instructional quality, students are strongly influenced by impressive situations that occurred a few days or a few weeks ago. However, in the research of Wagner et al., (2016) the ratings from teachers and students from single measurement points were reasonable approximations of time-consistent rating components.
According to Biemans, Jongmans, de Jong, and Bergen (1999) many teachers are not able to realistically assess their own instructional behavior. This might be caused by the fact that teachers tend to work in isolation and cannot compare their behavior with that of colleagues, while students can compare teachers. Den Brok et al., (2006) argue that wishful thinking and self-protection against disappointment could result in ratings that differ from those provided by students.

4.2.2 External raters’ scores compared to teachers’ perceptions
Research focusing solely on the perceptions of teachers and external raters is scarce. In a study by the Dutch Inspectorate of Education (2013), teachers were asked to rate their own instruction on 9 items, such as items about clear instruction and feedback. These perceptions were compared to inspectors’ ratings. Two-third of the teachers was more positive about their own teaching than the inspectors were.

4.2.3 External raters’ scores compared to students’ perceptions
Multiple studies report on the agreement between external raters’ scores and students’ perceptions of teaching. All studies show low to moderate agreement between the two, but this differs across scales. Maulana and Helms-Lorenz (2016) observed pre-service teachers with the International Comparative Analysis of Learning and Teaching lesson observation instrument (ICALT; van de Grift, 2007), and measured students’ general perceptions using a questionnaire based on the same constructs included in ICALT. Overall, students were more positive than external raters were. Students thought that the majority of the teachers had a high teaching level, whereas external raters held the opinion that most teachers had a low teaching level. This differed per scale: external raters were more positive than students about the learning climate, classroom management, and the clarity of instruction, while students were more positive about teachers’ activating learning, adaption of teaching, and teaching strategies. An explanation of these findings might be that it is hard for students to recognize the quality of more complex teaching behaviors (Maulana & Helms-Lorenz, 2016). Downer, Stuhlman, Schweig, Martínez, and Ruzek (2015) investigated the extent to which individual or classroom-level student perceptions about classroom interactions by upper elementary-aged students align with external raters’ judgements. Teachers were observed with the Classroom Assessment Scoring System (CLASS; Pianta, Hamre, & Mintz, 2012) and students filled in a questionnaire for rating the same three domains as in CLASS, the CLASS-Student-Report. Student reports concerning emotional support and classroom organization were consistently associated with external raters’ scores, whereas instructional support was not.
Several studies report correlations between external raters’ scores and students’ perceptions. De Jong and Westerhof (2001) measured the correlation between external raters’ scores and students’ perceptions of the extent to which teachers teach using a direct instruction style. Correlations were low. Similar correlations were found in the Measures of Effective Teaching (MET) project (Bill & Melinda Gates Foundation, 2010) by Ferguson and Danielson (2014), who compared external raters’ scores based on the Framework for Teaching (FfT; Danielson, 2013) and students’ perceptions of lesson quality based on the Tripod 7Cs model (Tripod Education Partners, 2016) in grades 4 through 8. Low to moderate correlations were found by van der Scheer, Bijlsma, & Glas (n.d.) in a study in fourth grade. Teachers were observed and rated with the ICALT lesson observation instrument and the student questionnaire used was partly based on the Tripod 7C’s model. No clear pattern could be identified in the correlations, the highest correlation was found between two scales that were not related content wise. In the research by Ellis, Malloy, Meece, and Sylvester (2007), it was explored whether students and external raters agreed on whether classroom practices were reform-oriented or not. Teachers were observed with the Reformed Teaching Observation Protocol (RTOP), and general student perceptions were measured using a questionnaire based on the RTOP. The authors found moderately strong correlations between the two measures on three of the four scales.

Several issues may have contributed to the low correlations that were found in the studies that we referred to in the previous. First, the observation instruments and the student perception measures in these studies are presumed to measure the same theoretical constructs, however the instruments all differ in their operationalization at the item level (De Jong & Westerhof, 2001; Maulana & Helms-Lorenz, 2016).

Second, in the studies it is assumed that the observation scores for one or a few lessons as given by external raters can be generalized to a broader period (e.g., a whole school year), the period that students also keep in mind when providing their general perceptions of a teacher’s teaching quality. However, often no evidence was found for this assumption (for example, in a generalizability study). Observation studies investigating the generalization of observation scores showed that in order to obtain reliable observation scores, multiple external raters need to observe multiple lessons (Hill et al., 2012; Nava et al., 2018; van der Lans, van de Grift, van Veen, & Fokkens-Bruinsma, 2016). In most of the studies that were mentioned, the researchers did not work with multiple external raters and the number of observations ranged from one to three. Research by van der Lans (2018) showed that conducting multiple observations
Quality lies in the eye of the beholder by multiple external raters, not only increases the reliability of the ratings but also had an effect on the correlations between external raters’ scores and students’ perceptions of the quality of the lessons. The expected correlation approximately doubles if a lesson is rated by three external raters instead of one rater visiting a single lesson.

Third, problems may occur if student perceptions and external lesson observations are not conducted at the same moment in time. For example, in the research by De Jong and Westerhof (2001), three observations were conducted during the whole school year, while students’ perceptions were measured at the end of the year. The scores resulting from both approaches are presumed to be generalizable to the whole school year, although student perceptions can vary considerably across the school year; they for example tend to be lower at the end of the school year (van der Scheer, 2016; van Geel & Keuning, 2016). In the study by van der Scheer et al. (n.d.), it was assumed that teaching quality is stable across school years. In that study, the external observations were not conducted during the same school year as the school year in which the student perceptions were measured, while research showed that the year-to-year stability of observation scores is only low to moderate, lower than within-year estimates because the students are different (Polikoff, 2015).

4.2.4 External raters’ scores compared to students’ and teachers’ perceptions
We only found two studies that compared external raters’ scores to students’ and teachers’ perceptions of teaching in the same study. Also, in these studies, the agreement between the different types of external raters was low and similar potential explanations as the ones that were presented in the previous section might play a role. Clausen (2002) used data from the TIMSS (Third International Mathematics and Science Study) in Germany, to compare the three methods of measuring teaching quality. Students filled in a questionnaire including 22 scales about teaching (e.g., the pacing of teaching and student participation), and based on 12 of these scales, a teacher self-report and an observation instrument were developed. Items for all three rater groups were similar, but not identical. The results showed that overall agreement between the three rater groups was low, however the external raters and teachers/students agreed more on aspects that were easy to observe according to Clausen (2002), such as student discipline.

Lawrenz et al. (2003) investigated if teachers, students, and external raters agreed on the forms of science instruction (e.g., class discussion or a demonstration) that occurred the most in classrooms. External raters observed grade nine science
classrooms four times over a period of two years using an observation instrument in which the most common forms of instruction are documented every 5-minutes. Every school year, students and teachers filled in a questionnaire with items that paralleled the classroom activities in the observation form. The results were mixed, but overall there were no or low correlations between the three groups. According to Lawrenz et al. (2003), the results showed that the three measures of teaching quality are measuring distinct things.

4.2.5 The current study
In this study, we focus on the comparison of external raters’ scores, and students’ and teachers’ perceptions of the same lessons, since these three methods for measuring teaching quality have rarely been compared within a single study. Our research design also differs from the studies referred to previously on multiple aspects. In our study, students’ and teachers’ perceptions of teaching quality were measured at the end of each of three lessons. All lessons were video-taped, and three external raters reviewed all the videotaped lessons and rated them after watching the whole lesson. The items used were similar for all three rater groups and only concerned the rating of that single lesson and not a general rating of a teacher’s teaching quality. Students’ perceptions were measured by means of a smartphone application (the Impact! tool) that students can use to provide feedback to the teacher at the end of a lesson. The application includes 15 closed questions about the quality of teaching in that lesson. Teacher perceptions were measured by means of a questionnaire that included 15 items that were as similar as possible to the student items. The external raters used two classroom observation systems, one that included 15 items that were as similar as possible to the student items, called the Impact! observation instrument. Since the first data about the reliability and validity of the scores on this observation system were gathered in this study, we chose to also use a thoroughly researched classroom observation system. We searched for an system that was based on similar literature as the Impact! items and chose to use the ICALT (van de Grift, 2007). Finally, in contrast with most of the studies that were discussed here, we used a combination of Item Response Theory, Generalizability Theory, and Multilevel Models to analyze the data.

The research question for this study is: How do external raters’ scores of teaching quality relate to students’ and teachers’ perceptions of teaching quality? We expected to find higher correlations between the three groups of raters than in previous research, since our research design used an equivalent reference period (a single lesson) for
students, teachers, and external raters, the content of the items were the same for all three groups of raters, and the three lessons per teacher were observed by three trained external raters. Based on the previous studies, we also hypothesized that teachers’ perceptions would be more positive than students’ and external raters’ scores, and that external raters would give the lowest scores of all three rater groups.

4.3 Method

4.3.1 Design and procedures
The teachers and students in this study all were in the experimental group of an intervention study into the effect of the Impact! tool (Bijlsma et al., n.d.). The intervention took place in school year 2016/2017 and lasted four months, during which teachers could use the Impact! tool as much as they wanted to. In this period, we collected data during three lessons from each participating teacher. Whenever possible, the three lessons were scheduled one month apart from each other in agreement with the teachers. This was not always possible due to tests and other practical circumstances.

In all three lessons, students were asked to use the Impact! tool at the end of the lesson and teachers were asked to fill in the questionnaire. All three lessons were videotaped and were rated later by three trained external raters.

4.3.2 Participants

Teachers
Out of the 28 teachers in the experimental group in the intervention study, 25 participated in this study. Three teachers did not participate, because they did not want to be videotaped. All teachers were mathematics teachers in the ninth grade of HAVO (a secondary school type in the Netherlands) and participation was voluntary.

The teachers (60% male) were on average 40.68 years old ($SD = 10.80$ years) and had on average 10.46 years of experience in teaching mathematics in secondary education ($SD = 8.32$ years). A group of three teachers and a teacher pair worked at the same school, all other teachers worked at different schools in the Netherlands (22 different schools).

Students
A total of 608 students participated in this study by providing their perceived lesson quality at least once. Students (51.3% female) were almost all 14 or 15 years old. Class
sizes ranged from 10 to 30 students per class, with a mean of 24 students ($SD = 5$ students). The response rate per lesson differed for multiple reasons, for example, because some students were not present. If students’ smartphones did not work properly, then they could fill in the questions on paper. On average $78.2\%$ ($SD = 12.3\%$) of the students responded. Student data are missing for one lesson due to a temporary malfunction of the Impact! tool.

**External raters**

All videotaped lessons were observed by three external raters. All external raters were (former) teachers with a Master’s degree in Education. The external raters were trained for using the two classroom observation systems (the Impact! observation instrument and ICALT). They first followed the ICALT training (5 hours) and obtained the ICALT rater certification. Afterwards, the external raters were trained in the Impact! observation instrument (8 hours). Since the external raters had to use both observation instruments during a single observation, the external raters were additionally trained for another 15 hours in using both instruments at the same time by rating videotaped lessons and discussing the ratings. External raters had to meet an inter-rater reliability standard after the training for both observation instruments (minimum ICC(2, 3) of .70).

**4.3.3 Instruments**

**Student perception measure**

Student perceptions were measured by means of the Impact! tool (called the IMPACT! app at the time of the research), a smartphone application developed by the University of Twente to gather student feedback for teachers. Teachers could use the application to gather feedback on a single lesson. If teachers chose to use the application, then students filled in 15 items on a 4-point scale (totally disagree to totally agree) at the end of that lesson. The items measured a student’s perception of the lesson, for example, whether the student understood the explanation by the teacher and whether he/she felt safe during the lesson (see Appendix E for all items). All items covered a number of scientifically determined characteristics of effective lessons. Students also answered an additional open question on how the teacher could improve the lesson. After the lesson, the teachers could see students’ responses in a web-environment. For more information on the Impact! tool, its development, and research into validity of the Impact! scores, (see Bijlsma, Glas, & Visscher, 2018a, 2018b; Bijlsma et al., n.d.).
Quality lies in the eye of the beholder

Teacher perception measure

The perception of teachers was measured with a paper-and-pencil questionnaire that included the same 15 closed items as the Impact! tool for students. The student items had been reformulated to make them suitable for teachers. For example, the student item ‘The teacher provided enough time to work on the assignments during the lesson’ had been reformulated into the following item for teachers: ‘I provided students enough time to work on the assignments during the lesson’. See Appendix E for all items. All items were scored on a 4-point scale (totally disagree to totally agree).

Classroom observation systems

The trained external raters used two classroom observation systems to rate all 75 lessons (3 lessons per teacher). The Impact! observation instrument included the same 15 closed items as the Impact! tool for the students and as the teacher questionnaire. However, the Impact! tool items had been reformulated to make them suitable for use by external raters; for example the student item ‘The teacher ensured that I actively participated during the lesson’ had been reformulated into ‘The teacher ensured the active participation of the students during the lesson’. See Appendix E for all external rater items. All items were scored on a 4-point scale (totally disagree to totally agree). For the use of the Impact! observation instrument, a rater manual had been developed, which included scoring rules for each item that indicated when to assign a specific score. A first version of the scoring rules was developed in a focus group with four didactical experts of the University of Twente, in which they indicated which teacher and student behavior would fit the four score-points. The first version was revised based on the input of three other (didactical) experts in the field.

The second classroom observation system that was used was the ICALT, which is a frequently researched and used instrument (e.g., van de Grift, 2007; van de Grift, 2009; van de Grift, 2014). The ICALT instrument measures teaching quality with 35 items across 7 scales (safe and stimulating learning climate, efficient classroom management, clear instruction, activating teaching methods, adaptation of instruction, teaching learning strategies, and engagement of students). The items were scored on a 4-point Likert scale (predominantly weak to predominantly strong). The ICALT does not provide scoring rules that specify when to assign a specific score (e.g., whether the observed quality equates a score 1 or 2). During the additional training, the external raters proved to be in need of such scoring rules. ICALT scoring rules were therefore
developed for this project, based on the information that had been provided during the official ICALT training.

4.3.4 Analyses

The model

The model used to analyze the item responses of the students, the teachers, and the external raters on the Impact! tool, and the item responses of the external raters on the ICALT instrument is a combination of an item response theory (IRT) model (Lord, 1980) and generalizability theory (GT) model (Brennan, 2001). The IRT model combines the item responses on the 4-point Likert scale into one measure. This measure is a so-called latent variable, say, a teacher proficiency variable. GT models are useful for disentangling the influence of factors such as raters, tasks, and time points on measurements and the definition of coefficients of agreement and reliability in complex measurement situations. The IRT model and the GT models involved in the present analyses will be discussed in turn.

To define the IRT model used here, we label the items as \( k = 1, \ldots, K \), with \( K = 15 \) for the Impact! tool and \( K = 35 \) for the ICALT observation instrument. The response to item \( k \) will be denoted by \( U_k = m \), with \( m = 0, \ldots, 3 \) for the four points on the 4-point Likert scale. The IRT model used is the generalized partial credit model (GPCM model) by Muraki (1992). In this model, the probability of a response in category \( m \) is given by

\[
p(U_k = m \mid \theta) = \Psi_{km}(\theta) = \frac{\exp\left(m\alpha_k\theta - \sum_{n=1}^{m} \delta_{kn}\right)}{1 + \sum_{h=1}^{3} \exp\left(h\alpha_k\theta - \sum_{n=1}^{h} \delta_{kn}\right)}.
\]  

In this formula, \( \theta \) is the latent variable to be measured and \( \alpha_k \) and \( \delta_{km} (m = 1, \ldots, 3) \) are the item parameters of item \( k \). The item parameter \( \alpha_k \) is called the discrimination parameter because it gauges the dependence of the responses on the latent variable \( \theta \). Further, \( \Psi_{km}(\theta) \) is a so-called item response curve, that is, a curve over the domain of \( \theta \) (the real numbers), and together with the discrimination parameter, the so-called location parameters \( \delta_{km} (m = 1, \ldots, 3) \), define the form of the item response curve (for more information on the model, see Muraki, 1992). The item parameters can also be used to estimate the contribution of an item on the measurement precision locally at \( \theta \). It will be assumed that the item parameters of the Impact! tool are
constant over students, raters, and teachers. This assumption will be tested below. Next, we introduce three GT models, for the raters, the students, and the teachers, respectively.

**External raters**

For the external raters, let \( \Theta_{1jrt} \) be the measurement of teaching quality of teacher \( j \) by external rater \( r \) in lesson \( t \). The first subscript of \( \Theta \), that is, the “1”, indicates that this is the proficiency of teachers as perceived by external raters. The probability of a response to item \( k \) of an external rater \( r \) (\( r = 1, \ldots, R, R = 3 \)) regarding lesson \( t \) (\( t = 1, \ldots, T, T = 3 \)), follows the GPCM defined above in Formula 1, that is, with the proper substitution of indices, the probability is given by

\[
p(U_{1jrtk} = m | \Theta_{1jrt}) = \Psi_{km}(\Theta_{1jrt})
\]

To model the variability over external raters and lessons, the measurement \( \Theta_{1jrt} \) is decomposed as

\[
\Theta_{1jrt} = \Theta_{1j} + \tau_{1r} + \tau_{1t} + \tau_{1jr} + \tau_{1jt} + \varepsilon_{1jrt}
\]

where the terms in the decomposition are defined in Table 4.1. The effects are assumed to be random and normally distributed.

### Table 4.1 Factors in generalizability model for the external raters

<table>
<thead>
<tr>
<th>Effect</th>
<th>Interpretation</th>
<th>Variance Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Theta_{1j} )</td>
<td>Main Effect Teachers</td>
<td>( \sigma_{1j}^2 )</td>
</tr>
<tr>
<td>( \tau_{1r} )</td>
<td>Main Effect External Rater</td>
<td>( \sigma_{1r}^2 )</td>
</tr>
<tr>
<td>( \tau_{1t} )</td>
<td>Main Effect Lessons</td>
<td>( \sigma_{1t}^2 )</td>
</tr>
<tr>
<td>( \tau_{1jr} )</td>
<td>Interaction Teachers and External Raters</td>
<td>( \sigma_{1jr}^2 )</td>
</tr>
<tr>
<td>( \tau_{1jt} )</td>
<td>Interaction Teachers and Lessons</td>
<td>( \sigma_{1jt}^2 )</td>
</tr>
<tr>
<td>( \tau_{1rt} )</td>
<td>Interaction Effect External Raters and Lessons</td>
<td>( \sigma_{1rt}^2 )</td>
</tr>
<tr>
<td>( \varepsilon_{1jrt} )</td>
<td>Error Component</td>
<td>( \sigma_{1jrt}^2 )</td>
</tr>
</tbody>
</table>

Since a teacher’s proficiency is the main objective of the measurement, the main effect of the teachers, \( \Theta_{1j} \), is the measure of interest and the other factors are a
nuisance. This notion is formalized in the definition of two so-called generalizability coefficients: agreement and reliability. The coefficient of agreement is defined as

\[ \rho_A = \frac{\sigma_{11}^2}{\sigma_{11}^2 + \sigma_{1r}^2 / R + \sigma_{1t}^2 / T + \sigma_{1jrt}^2 / RT + \sigma_{1jrt}^2 / RT} \]  

(3)

while the reliability coefficient is defined as

\[ \rho_R = \frac{\sigma_{11}^2}{\sigma_{11}^2 + \sigma_{1jt}^2 / R + \sigma_{1jt}^2 / T + \sigma_{1jrt}^2 / RT} \]  

(4)

Agreement is relevant for absolute judgements (for instance, related to some absolute standard of proficiency that must be met), while reliability is relevant for relative judgements (such as ordering the proficiency of the teachers). In the latter case, the variance of external raters and lessons and their interaction disappears from the denominator, because the measures are averaged over raters and lessons.

Students

For the students, the model can be developed roughly along the same lines; however, as will become clear, the fact that students’ observations are nested within teachers leads to some restrictions. Let \( \theta_{2ijt} \) be the measurement of teaching quality of teacher \( j \) by student \( i \) in lesson \( t \). The first subscript of \( \theta \), that is, the “2”, indicates that this is the proficiency as perceived by students. The probability of a response to item \( k \) of a student \( i \) in lesson \( t \) again follows the GPCM as defined above in Formula 1; that is, with the proper substitution of indices, the probability is given by

\[ p(U_{2ijk} = m \mid \theta_{2ijt}) = \Psi_{km}(\theta_{2ijt}) \]  

To model variability over students and lessons, the measurement is decomposed as

\[ \theta_{2ijt} = \theta_{2j} + \omega_{2i} + \omega_{2ij} + \omega_{2jt} + \epsilon_{2ijt} \]  

(5)

where the terms in the decomposition are defined in Table 4.2.
Table 4.2 Factors in generalizability model for students

<table>
<thead>
<tr>
<th>Effect</th>
<th>Interpretation</th>
<th>Variance Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_{2j}$</td>
<td>Main Effect Teachers</td>
<td>$\sigma_{22}^2$</td>
</tr>
<tr>
<td>$\omega_{2t}$</td>
<td>Main Effect Lessons</td>
<td>$\sigma_{2t}^2$</td>
</tr>
<tr>
<td>$\omega_{2i\mid j}$</td>
<td>Main Effect Students</td>
<td>$\sigma_{2i\mid j}^2$</td>
</tr>
<tr>
<td>$\omega_{2j\mid t}$</td>
<td>Interaction Teachers and Lessons</td>
<td>$\sigma_{2j\mid t}^2$</td>
</tr>
<tr>
<td>$\epsilon_{2i\mid t\mid j}$</td>
<td>Error Component (confounded with interaction students and lessons conditional on teachers)</td>
<td>$\sigma_{2i\mid t\mid j}^2$</td>
</tr>
</tbody>
</table>

Since the students are nested within the teacher, the between teacher variance is confounded with the between-student-groups variance, and since students are not crossed with teachers, agreement is not an issue here. However, for the same reason, variance of students within classes does become an issue, because each teacher is rated by a different group of students. Therefore, the reliability coefficient is defined by

$$\rho_R = \frac{\sigma_{22}^2}{\sigma_{22}^2 + \frac{\sigma_{2t}^2}{T} + \frac{\sigma_{2i\mid j}^2}{T} + \frac{\sigma_{2i\mid t\mid j}^2}{\tilde{N}} + \frac{\sigma_{2i\mid t\mid j}^2}{NT}},$$

where $T$ stands for a weighted mean number of time points over students and $\tilde{N}$ stands for the average number of students rating a teacher.

**Teachers**

A response of a teacher on the Impact! tool is denoted as $U_{3jk}$ and the response is again modelled by the GPCM, so $p(U_{3jk} = m \mid \theta_{3jk}) = \Psi_{km}(\theta_{3jk})$, where $\theta_{3jk}$ is assumed to be normally distributed, that is

$$\theta_{3jk} = \theta_{3j} + \epsilon_{jt} \quad \text{with} \quad \epsilon_{jt} \sim N(0, \sigma^2).$$

Since the teachers only assess themselves, agreement and reliability cannot be estimated here.
Chapter 4

» Relationship between the three groups of raters

The research question defined above relates to the relationship between the three different groups of raters. To estimate this relationship, the model is enhanced with the assumption that the proficiency measures by raters, students and teachers, that is, the values of $\theta_{1j}$, $\theta_{2j}$, and $\theta_{3j}$ have a multivariate normal distribution, that is,

$$
\begin{bmatrix}
\theta_{1j} \\
\theta_{2j} \\
\theta_{3j}
\end{bmatrix} \sim N
\begin{bmatrix}
\mu_1 \\
\mu_2 \\
\mu_3
\end{bmatrix},
\begin{bmatrix}
\sigma_{11} & \sigma_{12} & \sigma_{13} \\
\sigma_{21} & \sigma_{22} & \sigma_{23} \\
\sigma_{31} & \sigma_{32} & \sigma_{33}
\end{bmatrix}.
$$

The covariance matrix serves as a measure of the dependence between the three rater groups (external raters, students, and teachers). Further, since the item parameters are constant over the three rater groups, the three means $\mu_1$, $\mu_2$, and $\mu_3$ give the average level of the rater groups on a common scale. Thus, the higher a mean, the higher the average rating obtained by that group of raters.

Estimation procedures

The models were estimated in a Bayesian framework. The main motivation for this choice is the availability of software that supports building unique and complex models such as the model developed for the present application. The programs used were OpenBugs (Lunn, Spiegelhalter, Thomas, & Best, 2009) and Lexter (Glas, 2017). OpenBugs was used to estimate the model, while Lexter was used to compute model fit indices based on the estimates issued by OpenBugs.

OpenBugs is one among several open source packages that provide a general purpose Markov chain Monte Carlo (MCMC) sampler to estimate complex IRT models in a Bayesian framework. The script for estimation of the overall model is provided in Appendix F.

Lexter is a publicly available package for estimation of IRT models and evaluation of their model fit both in a likelihood-based framework and in a Bayesian framework.
4.4 Results

Before we can answer our main research question, we first need to answer the following two questions:

- *Do the teaching quality items for all three rater-groups form IRT scales?*
- *Do the teaching quality items function similarly across the three rater-groups?*

4.4.1 Do the teaching quality items for all three rater-groups form IRT scales?

**Item parameters**

To answer this question, IRT analyses were conducted to investigate whether the items for all three groups fit an IRT model. Appendix G contains the item parameter estimates and information values for each of the groups. These absolute values of the estimates for the three rater-groups are not on a common scale yet, and therefore for a first indication of the comparability of the three groups, we provide in Table 4.3 the correlations between the item parameter estimates. We see that the three scales functioned quite similarly for the three groups: the correlations of the item parameters estimates are high between the students and teachers, and between the external raters and students. The correlations are moderate between the external raters and teachers.

<table>
<thead>
<tr>
<th>Assessors</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>External raters / Students</td>
<td>0.70</td>
</tr>
<tr>
<td>External raters / Teachers</td>
<td>0.61</td>
</tr>
<tr>
<td>Students / Teachers</td>
<td>0.85</td>
</tr>
</tbody>
</table>

**Item fit**

Using these item parameter estimates, item fit was investigated for all three rater-groups. To test the model for external raters, the 225 scores provided by three external raters on 3 lessons of 25 teachers were partitioned into low, medium, and high ratings. After partitioning the scores, the observed average item scores and their posterior expectation given the parameter estimates were computed. Table 4.4 presents the results. The last column gives the average absolute difference between the observed and the expected item scores. Large absolute differences, larger than 0.10, indicate a lack of model fit for the items (Glas, 1998, 1999, 2005). Inspecting Table 4.4, we can conclude that the model fits for all but one item. For the item on
summarizing lesson content, the absolute difference is slightly above 0.10. For this item, the observed score given low or high ratings is higher than expected while the observed score given medium ratings is lower than expected.

<table>
<thead>
<tr>
<th>Item</th>
<th>Low ratings</th>
<th>Medium ratings</th>
<th>High ratings</th>
<th>Mean absolute difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Obs.</td>
<td>Exp.</td>
<td>Obs.</td>
<td>Exp.</td>
</tr>
<tr>
<td>Learning goals</td>
<td>0.97</td>
<td>0.92</td>
<td>1.07</td>
<td>1.16</td>
</tr>
<tr>
<td>Explanation</td>
<td>1.53</td>
<td>1.54</td>
<td>2.01</td>
<td>1.99</td>
</tr>
<tr>
<td>Prior knowledge</td>
<td>1.89</td>
<td>1.90</td>
<td>2.10</td>
<td>2.12</td>
</tr>
<tr>
<td>Checking for understanding</td>
<td>1.22</td>
<td>1.27</td>
<td>1.89</td>
<td>1.82</td>
</tr>
<tr>
<td>Asking questions</td>
<td>1.14</td>
<td>1.11</td>
<td>1.72</td>
<td>1.75</td>
</tr>
<tr>
<td>Additional explanation</td>
<td>1.57</td>
<td>1.63</td>
<td>2.07</td>
<td>2.01</td>
</tr>
<tr>
<td>Classroom work climate</td>
<td>1.30</td>
<td>1.29</td>
<td>1.80</td>
<td>1.83</td>
</tr>
<tr>
<td>Safe climate</td>
<td>1.64</td>
<td>1.66</td>
<td>2.01</td>
<td>1.97</td>
</tr>
<tr>
<td>Pleasant climate</td>
<td>1.73</td>
<td>1.68</td>
<td>1.91</td>
<td>1.99</td>
</tr>
<tr>
<td>Lesson pace</td>
<td>1.57</td>
<td>1.58</td>
<td>1.91</td>
<td>1.88</td>
</tr>
<tr>
<td>Feedback</td>
<td>1.51</td>
<td>1.51</td>
<td>1.84</td>
<td>1.90</td>
</tr>
<tr>
<td>Active participation</td>
<td>1.24</td>
<td>1.21</td>
<td>1.78</td>
<td>1.79</td>
</tr>
<tr>
<td>Time for assignments</td>
<td>1.73</td>
<td>1.68</td>
<td>1.96</td>
<td>2.00</td>
</tr>
<tr>
<td>Summarizing lesson content</td>
<td>0.25</td>
<td>0.16</td>
<td>0.15</td>
<td>0.32</td>
</tr>
<tr>
<td>Ability to complete assignments</td>
<td>1.68</td>
<td>1.62</td>
<td>1.88</td>
<td>1.95</td>
</tr>
</tbody>
</table>
The tables 4.5 and 4.6 provide analogous results for the students and teachers. We note that for the students, the model fits for all items with absolute differences for all items being smaller than the 0.10 norm, even below 0.04. For teachers, the mean absolute difference for three items slightly exceeds the norm, namely the items on time for assignments, classroom work climate and, again, about summarizing lesson content.

The overall conclusion is that, though not all items fit perfectly, the scales function well for the three groups of raters.

<table>
<thead>
<tr>
<th>Item</th>
<th>Low ratings</th>
<th></th>
<th>Medium ratings</th>
<th></th>
<th>High ratings</th>
<th></th>
<th>Mean absolute difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning goals</td>
<td>1.74</td>
<td>1.72</td>
<td>2.08</td>
<td>2.14</td>
<td>2.60</td>
<td>2.57</td>
<td>0.03</td>
</tr>
<tr>
<td>Explanation</td>
<td>1.69</td>
<td>1.69</td>
<td>2.15</td>
<td>2.16</td>
<td>2.65</td>
<td>2.64</td>
<td>0.00</td>
</tr>
<tr>
<td>Prior knowledge</td>
<td>1.63</td>
<td>1.59</td>
<td>2.01</td>
<td>2.01</td>
<td>2.44</td>
<td>2.47</td>
<td>0.02</td>
</tr>
<tr>
<td>Checking for understanding</td>
<td>1.33</td>
<td>1.37</td>
<td>1.87</td>
<td>1.86</td>
<td>2.44</td>
<td>2.42</td>
<td>0.02</td>
</tr>
<tr>
<td>Asking questions</td>
<td>1.31</td>
<td>1.35</td>
<td>1.79</td>
<td>1.76</td>
<td>2.30</td>
<td>2.29</td>
<td>0.03</td>
</tr>
<tr>
<td>Additional explanation</td>
<td>1.60</td>
<td>1.56</td>
<td>2.04</td>
<td>2.06</td>
<td>2.62</td>
<td>2.63</td>
<td>0.02</td>
</tr>
<tr>
<td>Classroom work climate</td>
<td>1.56</td>
<td>1.54</td>
<td>1.98</td>
<td>1.99</td>
<td>2.48</td>
<td>2.49</td>
<td>0.01</td>
</tr>
<tr>
<td>Safe climate</td>
<td>1.72</td>
<td>1.69</td>
<td>2.11</td>
<td>2.14</td>
<td>2.64</td>
<td>2.64</td>
<td>0.02</td>
</tr>
<tr>
<td>Pleasant climate</td>
<td>1.64</td>
<td>1.64</td>
<td>2.12</td>
<td>2.10</td>
<td>2.57</td>
<td>2.59</td>
<td>0.01</td>
</tr>
<tr>
<td>Lesson pace</td>
<td>1.70</td>
<td>1.69</td>
<td>2.12</td>
<td>2.13</td>
<td>2.58</td>
<td>2.59</td>
<td>0.01</td>
</tr>
<tr>
<td>Feedback</td>
<td>1.64</td>
<td>1.61</td>
<td>2.06</td>
<td>2.09</td>
<td>2.58</td>
<td>2.59</td>
<td>0.02</td>
</tr>
<tr>
<td>Active participation</td>
<td>1.50</td>
<td>1.50</td>
<td>1.96</td>
<td>1.98</td>
<td>2.53</td>
<td>2.51</td>
<td>0.02</td>
</tr>
<tr>
<td>Time for assignments</td>
<td>1.70</td>
<td>1.71</td>
<td>2.12</td>
<td>2.11</td>
<td>2.53</td>
<td>2.53</td>
<td>0.01</td>
</tr>
<tr>
<td>Summarizing lesson content</td>
<td>1.09</td>
<td>1.10</td>
<td>1.49</td>
<td>1.45</td>
<td>1.98</td>
<td>2.01</td>
<td>0.03</td>
</tr>
<tr>
<td>Ability to complete assignments</td>
<td>1.82</td>
<td>1.77</td>
<td>2.16</td>
<td>2.20</td>
<td>2.61</td>
<td>2.63</td>
<td>0.04</td>
</tr>
</tbody>
</table>
### Table 4.6 Item fit for the teachers

<table>
<thead>
<tr>
<th>Item</th>
<th>Low ratings</th>
<th>Medium ratings</th>
<th>High ratings</th>
<th>Mean absolute difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning goals</td>
<td>2.12</td>
<td>2.05</td>
<td>2.20</td>
<td>2.17</td>
</tr>
<tr>
<td>Explanation</td>
<td>1.86</td>
<td>1.87</td>
<td>2.00</td>
<td>2.03</td>
</tr>
<tr>
<td>Prior knowledge</td>
<td>2.09</td>
<td>2.03</td>
<td>2.08</td>
<td>2.18</td>
</tr>
<tr>
<td>Checking for understanding</td>
<td>2.00</td>
<td>1.98</td>
<td>2.07</td>
<td>2.08</td>
</tr>
<tr>
<td>Asking questions</td>
<td>1.71</td>
<td>1.71</td>
<td>1.83</td>
<td>1.83</td>
</tr>
<tr>
<td>Additional explanation</td>
<td>1.95</td>
<td>1.96</td>
<td>2.00</td>
<td>2.08</td>
</tr>
<tr>
<td>Classroom work climate</td>
<td>1.40</td>
<td>1.52</td>
<td>1.93</td>
<td>1.79</td>
</tr>
<tr>
<td>Safe climate</td>
<td>2.00</td>
<td>2.00</td>
<td>2.00</td>
<td>2.11</td>
</tr>
<tr>
<td>Pleasant climate</td>
<td>1.95</td>
<td>1.96</td>
<td>2.08</td>
<td>2.09</td>
</tr>
<tr>
<td>Lesson pace</td>
<td>1.91</td>
<td>1.79</td>
<td>1.93</td>
<td>1.98</td>
</tr>
<tr>
<td>Feedback</td>
<td>2.00</td>
<td>2.02</td>
<td>2.07</td>
<td>2.06</td>
</tr>
<tr>
<td>Active participation</td>
<td>1.75</td>
<td>1.71</td>
<td>1.89</td>
<td>1.92</td>
</tr>
<tr>
<td>Time for assignments</td>
<td>2.19</td>
<td>2.03</td>
<td>2.00</td>
<td>2.15</td>
</tr>
<tr>
<td>Summarizing lesson content</td>
<td>1.04</td>
<td>0.93</td>
<td>0.96</td>
<td>1.15</td>
</tr>
<tr>
<td>Ability to complete assignments</td>
<td>1.91</td>
<td>1.84</td>
<td>2.00</td>
<td>2.09</td>
</tr>
</tbody>
</table>

#### Generalizability coefficients

Table 4.7 gives the estimates of the generalizability coefficients. We use the more traditional terminology of standard errors and confidence regions, although the proper Bayesian terminology is posterior standard deviation and credibility region. The interpretations of these coefficients are similar. The coefficients in Table 4.7 are as defined in the formulas $3$, $4$, and $6$, respectively. A minimum coefficient of .70 is often mentioned as a norm for reliability when scores are used for low stake decisions (Evers et al., 2010). Notice that in Table 4.7, we see that the agreement of external raters is only moderate, while the reliability of external raters almost reaches .70. In the same table, we also see that the agreement between students is sufficient and the reliability good (above .80; Evers et al., 2010).
Quality lies in the eye of the beholder

Table 4.7 Generalizability coefficients for external raters and students

<table>
<thead>
<tr>
<th>Assessors</th>
<th>Coefficient</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>External raters</td>
<td>Agreement</td>
<td>.501</td>
</tr>
<tr>
<td></td>
<td>Reliability</td>
<td>.688</td>
</tr>
<tr>
<td>Students</td>
<td>Reliability</td>
<td>.822</td>
</tr>
</tbody>
</table>

4.4.2 Do the teaching quality items function similarly across the three rater-groups?

Using differential item functioning, we checked whether the items functioned similarly across all three groups of raters. The observed average item scores and their posterior expectation given the parameter estimates were computed. Table 4.8 provides insight into the differential item functioning for students compared to external raters. The last column provides a score that indicates whether items function differently between the two groups. Items function differently (show DIF) if this score exceeds a norm of 0.10 (Glas, 2005). The results show DIF for two items, the items regarding stating learning goals and the item on summarizing what students have learned. For both items, students rated their teachers higher than expected, whereas external raters rated them lower than expected.

Table 4.9 provides insight into the differential item functioning for teachers versus both raters and students together. The table shows DIF for 4 items. On three items, the items regarding prior knowledge, checking for understanding, and summarizing what was learned in the lesson, the teachers provided higher scores than expected. On summarizing lesson content and what was learned in the lesson, the external raters and students together rated teachers much lower than expected. On one item, classroom work climate, the teachers rated themselves lower than expected.

The overall conclusion is that it is reasonable to assume that most items in the scale function in the same way for all three rater-groups.
### Table 4.8 Differential item functioning students versus external raters

<table>
<thead>
<tr>
<th>Item</th>
<th>Low ratings</th>
<th>Medium ratings</th>
<th>DIF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Obs.</td>
<td>Exp.</td>
<td>Obs.</td>
</tr>
<tr>
<td>Learning goals</td>
<td>2.12 2.02</td>
<td>1.17 1.80</td>
<td></td>
</tr>
<tr>
<td>Explanation</td>
<td>2.14 2.15</td>
<td>1.97 1.93</td>
<td></td>
</tr>
<tr>
<td>Prior knowledge</td>
<td>2.01 2.06</td>
<td>2.19 1.87</td>
<td></td>
</tr>
<tr>
<td>Checking for understanding</td>
<td>1.87 1.89</td>
<td>1.79 1.64</td>
<td></td>
</tr>
<tr>
<td>Asking questions</td>
<td>1.80 1.82</td>
<td>1.72 1.59</td>
<td></td>
</tr>
<tr>
<td>Additional explanation</td>
<td>2.09 2.12</td>
<td>2.04 1.91</td>
<td></td>
</tr>
<tr>
<td>Classroom work climate</td>
<td>1.99 2.00</td>
<td>1.82 1.77</td>
<td></td>
</tr>
<tr>
<td>Safe climate</td>
<td>2.17 2.16</td>
<td>1.94 1.97</td>
<td></td>
</tr>
<tr>
<td>Pleasant climate</td>
<td>2.09 2.11</td>
<td>2.02 1.90</td>
<td></td>
</tr>
<tr>
<td>Lesson pace</td>
<td>2.11 2.10</td>
<td>1.84 1.90</td>
<td></td>
</tr>
<tr>
<td>Feedback</td>
<td>2.08 2.07</td>
<td>1.85 1.89</td>
<td></td>
</tr>
<tr>
<td>Active participation</td>
<td>1.98 1.98</td>
<td>1.77 1.74</td>
<td></td>
</tr>
<tr>
<td>Time for assignments</td>
<td>2.10 2.10</td>
<td>1.97 1.92</td>
<td></td>
</tr>
<tr>
<td>Summarizing lesson content</td>
<td>1.53 1.40</td>
<td>0.36 1.15</td>
<td></td>
</tr>
<tr>
<td>Ability to complete assignments</td>
<td>2.18 2.17</td>
<td>1.92 1.97</td>
<td></td>
</tr>
</tbody>
</table>
Table 4.9 Differential item functioning teachers versus raters and students

<table>
<thead>
<tr>
<th>Item</th>
<th>Low ratings</th>
<th>Medium ratings</th>
<th>DIF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Obs.</td>
<td>Exp.</td>
<td>Obs.</td>
</tr>
<tr>
<td>Learning goals</td>
<td>2.16</td>
<td>2.08</td>
<td>1.99</td>
</tr>
<tr>
<td>Explanation</td>
<td>2.16</td>
<td>2.20</td>
<td>2.12</td>
</tr>
<tr>
<td>Prior knowledge</td>
<td>2.28</td>
<td>2.11</td>
<td>2.03</td>
</tr>
<tr>
<td>Checking for understanding</td>
<td>2.16</td>
<td>1.95</td>
<td>1.86</td>
</tr>
<tr>
<td>Asking questions</td>
<td>1.87</td>
<td>1.85</td>
<td>1.79</td>
</tr>
<tr>
<td>Additional explanation</td>
<td>2.16</td>
<td>2.18</td>
<td>2.08</td>
</tr>
<tr>
<td>Classroom work climate</td>
<td>1.89</td>
<td>2.04</td>
<td>1.97</td>
</tr>
<tr>
<td>Safe climate</td>
<td>2.19</td>
<td>2.20</td>
<td>2.12</td>
</tr>
<tr>
<td>Pleasant climate</td>
<td>2.20</td>
<td>2.16</td>
<td>2.08</td>
</tr>
<tr>
<td>Lesson pace</td>
<td>2.07</td>
<td>2.15</td>
<td>2.08</td>
</tr>
<tr>
<td>Feedback</td>
<td>2.15</td>
<td>2.13</td>
<td>2.03</td>
</tr>
<tr>
<td>Active participation</td>
<td>2.00</td>
<td>2.03</td>
<td>1.95</td>
</tr>
<tr>
<td>Time for assignments</td>
<td>2.10</td>
<td>2.10</td>
<td>1.97</td>
</tr>
<tr>
<td>Summarizing lesson content</td>
<td>1.53</td>
<td>1.40</td>
<td>1.36</td>
</tr>
<tr>
<td>Ability to complete assignments</td>
<td>2.18</td>
<td>2.17</td>
<td>1.92</td>
</tr>
</tbody>
</table>
4.4.3 How do external raters’ scores of teaching quality relate to students’ and teachers’ perceptions of teaching quality?

After estimating the model separately for all three rater-groups, the data for the groups were pooled and a concurrent estimate was made of the complete model including the multivariate normal distribution displayed in Formula 8. We first report the results of the analyses in which we used external raters’ scores for the Impact! lesson observation instrument, second we will report the analyses with external raters’ scores on the ICALT observation instrument.

The estimates of the means, variances, and correlations between the three rater-groups scores of teachers’ proficiencies are given in Table 4.10. The raters were the least positive about teachers’ proficiencies, and scored them low on the latent scale compared to students and teachers. On average, teachers rated themselves highest. The mean ratings of the students had the highest variance. Table 4.10 further shows that there was no significant correlation between the scores given by external raters and the scores given by teachers, and between the scores from students and the ones from teachers. Only the correlation between the scores by external raters and the student scores differs significantly from zero. However, this correlation is rather low: .43.

<table>
<thead>
<tr>
<th>Rater Group</th>
<th>Impact! Scores</th>
<th>Correlation Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Variance</td>
</tr>
<tr>
<td>Raters</td>
<td>0.03</td>
<td>0.31</td>
</tr>
<tr>
<td>Students</td>
<td>0.56</td>
<td>0.76</td>
</tr>
<tr>
<td>Teachers</td>
<td>0.69</td>
<td>0.46</td>
</tr>
</tbody>
</table>

Note. ** p < 0.01

To assess the relation between the scores from external raters on the ICALT observation instrument and the scores for the Impact! instrument given by the external raters, students, and teachers, we made an additional concurrent Bayesian estimate. This estimate is similar to the previous one, however now the responses from the external raters for the ICALT instrument were included in the model as a whole.

Table 4.11 shows the correlations between the scores of the three rater-groups for the Impact! instruments and the external raters’ scores on ICALT observation...
Quality lies in the eye of the beholder

The correlation between the external raters on the Impact! observation instrument and the ICALT observation instrument is very high. The correlations between students’ and teachers’ Impact! scores and those by external raters on the ICALT observation instrument are similar to those presented in Table 4.10: a low correlation between external raters and students, and no correlation between the external raters and teachers.

<table>
<thead>
<tr>
<th>Table 4.11 Correlations between raters’ scores on ICALT and the scores by external raters, students, and teachers for the Impact! instruments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Rater Group</strong></td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td>Impact! scores</td>
</tr>
<tr>
<td>External raters</td>
</tr>
<tr>
<td>Students</td>
</tr>
<tr>
<td>Teachers</td>
</tr>
</tbody>
</table>

Note. ** p < 0.01

4.5 Conclusion and discussion

Measuring teaching quality is important to a broad array of educational stakeholders (Hill, Umland, Litke, & Kapitula, 2012), however there is debate about the best method for doing this (Chetty et al., 2011). In this study, we compared the results of three such methods: classroom observation by external raters and student and teacher surveys that measure perceptions of teaching quality. We answered the following research question in this study: How do external raters’ scores of teaching quality relate to students’ and teachers’ perceptions of teaching quality?

As hypothesized, teachers perceived the quality of their own teaching as the most positive, followed by the perceptions of students. External raters were much less positive than students and teachers. We expected to find higher correlations between the three methods than in previous studies, because, as opposed to previous studies, we:

- used very similar items for all three rater-groups, which we expected would lead to a similar interpretation of the items by all raters groups and therefore to higher correlations;
- used three measurement moments for all three rater-groups and worked with the same three external raters for each teacher, which we expected would lead to more reliable scores and therefore to higher correlations; and
- gathered data from all three rater-groups after the same lessons, and the items
only focused on that single lesson, thus, the reference period for all rater-groups was identical and shorter than the reference period in most other studies (in which the reference period often was several weeks, months, or a whole school year). We expected that this would also lead to higher correlations.

Despite incorporating these considerations into our design, we found no significant correlation between external raters’ scores and teachers’ perceptions of teaching quality, and no significant correlation between teachers’ and students’ perceptions. We found a fairly low correlation between the scores from external raters and students, higher than the correlations found in most studies (cf. van der Scheer et al., 2018).

4.5.1 Discussion of the results

The low agreement is probably not caused by a different interpretation of the items between rater-groups, as the Impact! items form a similar IRT scale for all three rater-groups.

The low agreement between the rater-groups may have been caused by the fairly low reliability of the scores given by the raters. Generally, lower reliabilities lead to the attenuation of correlations. In a D-study, in which we inserted four lessons or four raters in Formula 4, the reliability of the external raters rose to .81. The reliability of the scores given by the students was sufficient.

Although the relatively low reliability of the external raters’ scores could offer an explanation, it, in our view, does not completely explain the low agreement between the three rater-groups. Reflecting on each of the three comparisons leads to a number of other potential causes of no/low agreement between the rater-groups.

Students compared to teachers

Using the Impact! tool, students were asked whether the teacher indicated what they were going to learn, whether they understood teachers’ explanations of subject matter content, whether the teacher asked thought-provoking questions, whether they felt safe during the lesson, etc. The absence of a significant correlation between students’ and teachers’ perceptions might indicate that:

- teachers’ actions do not have the effect on students that teachers think. Thus, teachers and students disagreed about what teachers had accomplished on the dimensions measured by the Impact! items;
- teachers and students have different teaching quality standards or different
Quality lies in the eye of the beholder

expectations about the lesson based on their own personal preferences and experiences or because of the different roles and perspectives they have;

• students did not take answering the Impact! tool questions serious;
• teachers are less able to evaluate themselves critically (they on average provided higher scores than the students); and
• teachers’ self-ratings reflect in part wishful thinking, as suggested by den Brok et al. (2006).

External raters compared to students
A significant, but fairly low correlation was found between external raters’ ratings and students’ perceptions of teaching quality. This might indicate that:

• external raters and students have different standards, because external raters were trained for their rating task and used scoring rules, while students provided their, at least to some extent, subjective perceptions based on for example experiences with other teachers and personal preferences;
• it is hard for external raters to assess the effect of classroom teaching on students. For example, it may be difficult for an external rater to assess whether the explanation of the subject matter by the teachers was clear for the students or whether students felt safe in class; and
• students did not take answering the Impact! tool questions serious.

External raters compared to teachers
Furthermore, the external raters and teachers had different scores on the Impact! items (no significant correlation). This might indicate that:

• teachers have more knowledge of the context, their students, and their intentions in the lessons, and therefore rated the lessons differently compared to the external raters. For example, if the students work very hard compared to previous lessons, the teacher might be positive about their work attitude, while compared to other classes the external raters would still evaluate the student behavior as insufficient. This might have affected some of the scores of the Impact! items more than others;
• external raters and teachers have different standards. External raters were trained for their rating task and used scoring rules, while teachers provided their, at least partially, subjective perceptions based on for example personal preferences;
• teachers are less capable of evaluating themselves critically (they on average rated themselves higher than the external raters rated them); and
• teachers’ self-ratings reflect in part wishful thinking, as suggested by den Brok et al. (2006).
4.5.2 Implications
Since the correlations between the three rater-groups were either low or not significant, the results of the evaluation of teaching quality will differ substantially depending on which of the three rater-groups is used. Those who aim to measure teaching quality should keep this in mind and carefully select one or multiple methods. Often only one of the three methods is used in research and in practice. Since using multiple measures is costly and labor-intensive, an important question to ask is which perspective(s) is/are most important and most suitable for a specific context. It seems to us that in most situations multiple perspectives are relevant. For example, if teaching quality is measured for teacher professionalization or certification, all three perspectives seem to be relevant: external raters’ scores to assess to what extent teaching behavior meets specific (scientific) standards, students’ perceptions to know how the ‘clients’ perceive teaching quality (for example, whether they understood the explanation of subject matter content well), and teachers’ own perceptions to stimulate reflection (after the lesson, and after seeing the scores of the other rater-groups). Some rater-groups may be more able to rate specific aspects of teaching quality than others. For example, external raters may be more capable of evaluating teachers’ pedagogical content knowledge than students, teachers are the only rater-group aware of teachers’ intentions, and students know best whether they feel safe in the classroom or not.

Teachers rated their own teaching higher than students and external raters, and these scores did not correlate significantly with any of the other rater-groups. Therefore, it would seem important that teachers discuss the characteristics of their teaching with their students, and also have the opportunity to discuss their teaching quality with others, such as school leaders, peers, academic coaches, school counselors, and school inspectors. Furthermore, they should have the opportunity to observe other teachers in their classrooms.

4.5.3 Limitations and recommendations for future research
Since the scores of the three rater-groups turned out to differ substantially, an interesting follow-up question would be directed at which rater-group ratings correlate strongest with student results. However, since Dutch secondary schools do not use standardized tests in the ninth grade, we did not have the opportunity to collect such test scores from the participating students and therefore were unable to answer this question. This however seems a relevant question for future research (in grades in which standardized tests are used).
Quality lies in the eye of the beholder

We answered the research question after conducting a relatively small-scale research in a specific context (Dutch mathematics grade-nine classrooms) with specific items (the Impact! items). It is therefore relevant to further investigate the research question in other contexts (other subjects, different grades, different educational streams) and with different items (with different wording, that also measure other aspects of teaching quality than the items in the Impact! tool measure). We conducted the research in the months December until March. Since we know that student perceptions can differ across the school year (van der Scheer, 2016; van Geel & Keuning, 2016), this also may have affected the agreement between the rater-groups. Therefore, it would be interesting to conduct future research also at the beginning and at the end of the school year.

Finally, we have suggested that teachers should discuss the features of their teaching with their students and with external raters based on the ratings they receive from them. We wonder whether this will lead to greater agreement between the three rater-groups, this is another interesting topic for future research.
5 Conclusion and discussion
5.1 Introduction

A classroom observation system (COS) is used to systematically observe classroom practices and consists of (1) scoring tools (scales and items with a rating scale), (2) rating quality measures (procedures that promote accurate and reliable scoring by raters over time, such as rater training and a rater manual), and (3) sampling specifications (a description of the characteristics of the required sample of observations including the number of observations that should be conducted per teacher and the length of those observations; Bell, Dobbelaer, Klette, & Visscher, n.d.). This dissertation focuses on the quality and qualities of COSs. We noticed that on the one hand many researchers point to issues that should be taken into account when using COSs, whereas on the other hand those issues are often not (fully) addressed in educational practice and research. Since this brings the reliability and validity of COS scores into question (Coe, 2014), we decided to investigate these concerns in more depth. We therefore first reviewed the literature for the various kinds of quality criteria that COSs should meet, after which we searched for existing COSs and assessed to what extent the COSs meet these quality criteria. In the third study, we further investigated the validity of COSs by comparing COS scores with two other methods for measuring teaching quality: student and teacher surveys for measuring perceptions of teaching quality.

In this dissertation, we report on three studies. In this final chapter, we will first briefly summarize the three studies and reflect on the main findings for each study independently. After this, based on our findings, we will discuss some overarching conclusions and implications. We conclude this chapter with recommendations for future research.

5.2 Reflections on the main findings per study

5.2.1 The COS evaluation framework

In the first study, we developed a framework that brings together various issues to be taken into account when developing, selecting, or using a COS. We answered the question: Which quality requirements should a COS meet? We searched for quality criteria that COSs should meet in three strands of literature: the literature on COSs, the literature on testing and performance assessment (the Standards for Educational and Psychological Testing, AERA, APA, & NCME, 1999; the COTAN criteria for test quality, Evers, Lucassen, Meijer, & Sijtsma, 2010), and the literature on the argument-based approach to validity (Kane, 2006; 2013).
Chapter 5

The search resulted in the construction of a framework consisting of three parts. The first part is for evaluating the characteristics of the COS: the proposed use of the COS, its theoretical basis, the quality of the items, the standardization of the observations, and the measures for inter-rater reliability. The second part was developed for evaluating the reliable use of the COS: the availability of reliability information, an evaluation of the reliability information, and an evaluation of the research that provided this information. The third part of the framework is for evaluating the valid use of the COS scores by means of the argument-based approach to validity.

Looking back at the development process of the framework, we can state that the three strands of literature complement each other well in creating a comprehensive evaluation framework. Solely using the COSs literature would have been too limited for the purpose of our framework, because such publications are often very focused on one specific aspect of COSs (scores), for example the number of observations that need to be conducted. The general test and performance literature therefore was a useful addition, because it provides other more general but relevant criteria. Sometimes the criteria were too general, but based on the COS literature we were able to further specify the general criteria. We also used the COSs literature to add other criteria, for aspects of COSs that required more attention in our framework, than provided in the general literature, such as the training of raters. We chose to use the argument-based approach to validity, because this is a very useful approach to evaluate the validity of the scores by users for their own context, and it also is the preferred method for evaluating validity in the COS literature.

We believe the developed framework to be useful for (potential) COS users and developers, for obtaining insight on what should be taken into account when using classroom observation, and for evaluating the use of a COS in their own context. In our own research, the framework proved to be useful for developing a COS (which we did in the third study) and for evaluating existing COSs (which we did in the second study). In the latter study, COSs scored very differently on the framework criteria and thus the framework turned out to be a useful instrument for differentiating between COSs.

5.2.2 The quality of COSs

Using the framework that we developed in the first study, we conducted a review of the quality of COSs for measuring teaching quality in primary education. In doing so, we used the second study to answer the questions: Which COSs have been
developed to measure teaching quality in primary education, what is the quality of the COS materials and, what evidence is available regarding the reliability and validity of the scores these COSs produce? An extensive search resulted in 27 COSs that met our inclusion criteria (e.g., developed for primary education, publications about (research about) the COS after 1990 in English or in Dutch, research into the reliability and/or validity of the COS scores is available). All 27 COSs were reviewed by 2 reviewers, using the evaluation framework from the first study.

Based on the results of the review, it seems that of the three aspects in our definition of COSs, most attention is paid by instrument developers to the development of the scoring tools. The reviewers were quite positive about the quality of the COS materials in general. For example, for most COSs the documentation clearly specifies what it intends to measure, what the purpose of the COS is, and a description of the theory underlying the COS. The reviewers were generally less positive about the quality of the items in the COSs, as for example multiple behaviors were measured in a single item.

The instrument developers of all the included COSs paid at least some attention to rating quality measures (the second part of the COS definition). Most instrument developers developed scoring rules and mentioned the training of raters, however, for most COSs at least one of the reviewers doubted whether enough measures were taken to ensure that raters would score accurately and consistently. These doubts were caused, for example, by the fact that the rater training was short, or because no empirical evidence was available that raters could use the COS reliably. It is remarkable that for most COSs we did not find information about the follow-up training of raters, while it is known from COS literature that raters start to score differently after a while (i.e., rater drift; Casabianca, Lockwood, & McCaffrey, 2015; Doabler et al., 2015).

The reviewers were the least positive about the third part in our definition, the sampling specifications that COS developers provide. For all but one COS, the reviewers indicated that the instrument developers did not provide enough information about the sampling specifications. For example, only a few instrument developers were explicit about the number of lesson observations that should be conducted or when the lesson observations should be conducted during the day, week, and year.

The reviewers also evaluated the evidence available for the reliable and valid scoring
with the COSs. During the search for the observation systems, it already became clear that many of the scores resulting from COSs used in research and practice had no empirical backing for reliability and validity at all. For the 27 COSs included in the review, some evidence was available, however the results of the review showed that for most COSs, at least one of the reviewers was not convinced of the empirical backing of the scores.

Why would it be that the focus of COS developers is mostly on developing the scoring tools, and not on designing rating quality measures, specifying sampling specifications, and gathering empirical evidence for the reliability and validity of the scores? One reason could be that the scoring tools are developed first, and resources such as the time and money required to develop a COS that includes all parts of our definition and that is thoroughly researched are lacking. Another reason might be that instrument developers are not aware that rating quality measures, sampling specifications, and research are essential for a high quality COS. Observing teachers feels instinctively valid to many people (Coe, 2014), and therefore it might seem sufficient to developers to only develop sound scoring tools.

5.2.3 Classroom observation scores compared to students’ and teachers’ perceptions of teaching quality

In the third study, we conducted an empirical study into validity of COS scores, by comparing COS scores with students’ and teachers’ perceptions of teaching quality. We answered the question: How do raters’ scores of teaching quality relate to students’ and teachers’ perceptions of teaching quality? Answering this question contributes to our knowledge about the three methods for measuring teaching quality, and whether or not they provide the same information. This question has only been researched previously in a few studies.

In our study, we collected data from 25 grade nine mathematics classrooms. For three lessons per teacher, we collected students’ perceptions using a smartphone application (the Impact! tool), that students can use to provide feedback about the quality of a lesson to teachers at the end of that lesson. Teachers’ perceptions were also measured at the end of the same lessons, using a questionnaire. All lessons were videotaped, and three external raters rated all lessons afterwards as well. The Impact! tool included 15 items, that were reformulated for the external raters and the teachers.

The results showed that teachers’ perceptions of the quality of their own teaching
resulted in the highest scores of teaching quality, followed by students’ perceptions of teaching quality. External raters were much less positive than teachers and students. No significant correlation was found between external raters’ scores and teachers’ perceptions, nor between teachers’ and students’ perceptions. A fairly low correlation was found between external raters’ scores and students’ perceptions.

Multiple explanations for the low agreement between the three rater-groups were mentioned in chapter 4. One of the explanations is that it is hard for teachers and external raters to assess the effects of teaching on students, for example whether teachers’ questions made students really think or whether students felt safe in the classroom. Classroom observation by trained external raters is often seen as the golden standard, because it is conducted by an expert who assesses teaching quality according to scientific standards. Some aspects of teaching quality can probably be assessed best by external raters, such as whether the explanation of subject matter content by teachers is correct. However, it may be that the results found were caused by the fact that it is hard for external raters to validly evaluate the effects of teaching on students. Therefore, it seems important to include student perceptions in the evaluation of teaching quality, at least for some aspects of teaching quality.

5.3 General conclusions and implications

The studies in this dissertation all show the complexity of evaluating teaching quality validly using classroom observation (systems). The development of the framework showed that COSs need to meet many quality standards in order to obtain valid scores. Most COSs used in educational research and practice do not meet these quality standards, as it was shown by the review. The third study indicates that, even when COSs meet the quality standards, it is still questionable whether using only a COSs will result in valid scores of teaching quality, at least for some aspects of teaching quality, because the perceptions of students and teachers are also important to measure. It is therefore questionable whether many of the COSs used in educational research and practice can be validly used to evaluate teaching quality in research, to provide feedback to teachers, and to assess teachers in a summative manner.

This dissertation does not end with a plea for stopping the use of classroom observation. Evaluating teaching quality without conducting any classroom observation, and seeing what happens in the classroom, is hard to imagine (Coe, 2014). Also, as was stated previously, some aspects of teaching quality can probably be evaluated best
by means of classroom observation. What is necessary though, as was shown by the studies in this dissertation, is that those who aim to use classroom observation for measuring teaching quality need to consider carefully how they can obtain valid scores. We believe that there is an important role here for COS developers, COS users, governments, and other institutions; we will discuss this more in the following sections.

5.3.1 COS developers
We believe that COS developers can be the linking pin between what we know about classroom observation from research on the one hand and educational practice on the other hand. They can do so by developing COSs that meet the quality standards as described in our framework. This includes (1) developing COSs that include sound scoring tools, rating quality measures, and sampling procedures, (2) conducting research into the reliability and validity of the COS scores, and using this research to further develop the COS, and (3) providing guidelines to COS users on how to use a COS based on this research.

For COS developers to take on this role as a linking pin, they need to be aware of the importance of developing COSs that meet the quality standards, of the full complexity of COS development, and of the research required to provide valid COS scores to teachers.

5.3.2 COS users
COS users, for example educational scientists and practitioners, also need to become aware of the complexity of COS use. When users are aware of the importance of using COSs that meet the quality criteria in the framework, they will hopefully only choose COSs that meet these requirements or place higher demands on COS developers.

When COS users understand the importance of the guidelines within a COS, they will also be more willing to follow up the guidelines provided by COS developers. This will probably make the use of the COS more expensive and time consuming, for example, when multiple observations need to be conducted. The users need to understand that investing time and money in unreliable COS scores is a waste of (scarce) resources.

Developers have in our view the first responsibility to investigate the validity of the COS scores and develop scoring guidelines based on this research. However, it is
also important for users to evaluate and, if possible, to investigate COS use in their own context (for which they can also use the evaluation framework). The context in which a COS is used can have implications for its use, for example, if teachers are observed by raters they know such as their school leader. COS developers might not have developed guidelines for this, but it can still affect the validity of the COS scores. By using the argument-based approach, COS users might conclude that including an additional rater, that the teacher does not know is necessary, or the users might become aware of the fact that that they will not obtain valid scores for a specific purpose.

This awareness among COS users could be created by COS developers through training and rater manuals. Educational scientists who conduct research into COSs can also contribute to this by investing in that their results become known in educational practice.

5.3.3 Governments and other institutions

Awareness among COS developers and users of the importance of the quality aspects included in the framework is probably insufficient to change the field. In chapter 3, we already discussed how (local) governments can stimulate the development and use of COSs that meet the quality standards by (1) developing a database of such COSs, (2) develop an assessment system for COSs, and (3) provide independent assessments of COSs. When it is clear which COSs meet the quality requirements, governments can also require that schools only use such COSs. It is also important that governments provide appropriate resources to schools, since the development of COSs that meet the quality requirements will require more resources and therefore be more expensive. Furthermore, schools need the resources to follow the guidelines provided in these COSs, such as conducting multiple observations.

Educational scientists can also be stimulated to use COSs that meet the quality standards by other institutions. Institutions that reward research grants to educational researchers can require that such COSs are used in projects. The same kinds of requirements can be used by scientific journals, by including standards concerning COS use (or instrument use in general) in their (peer) review process.
5.4 Recommendations for future research

As was discussed previously, the review in chapter 3 shows that more research is needed to obtain knowledge of how the different COSs can provide valid scores for the different purposes that the COSs have been designed for. This is not only relevant for the developers and users of that specific COS, but can also provide general insights into the use of COSs, as for example is shown by the research into the MQI (e.g., Hill, Charalambous, & Kraft, 2012). The research currently available into COSs is often conducted in real educational practice (real teachers in real classrooms), however, the context is often not entirely authentic. For example, the raters are often part of the research team and have been specifically trained for the study. It would therefore be interesting to also investigate more how the COSs function in real educational settings.

Many COSs have been developed for the formative evaluation of teachers. An important aspect to investigate, which to our knowledge has not been researched sufficiently, is whether the COS scores provide a good basis for improvement. Based on the third study, it would be interesting to investigate this in combination with other methods for measuring teaching quality. Which method or combinations of methods provide the best insights enabling teachers to improve?

In this last chapter, we suggested that schools should also evaluate the use of a COS in their own context by using our evaluation framework. It is questionable whether schools are really able to do this. It would therefore be interesting to further investigate how schools can evaluate their own validity argument and which resources they need for this.
Conclusion and discussion
References
References
References


References


References


References


References


References


References


References


References


References


References


References


References


References


References


Appendices
**Appendix A: Evaluation framework**

<table>
<thead>
<tr>
<th>Evaluation framework</th>
<th>Part A. Evaluation of the relevant COTAN criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1. Theoretical basis of the COS</td>
</tr>
<tr>
<td></td>
<td>1.1 Is the purpose of the COS specified?</td>
</tr>
<tr>
<td></td>
<td>a. Are the constructs that the COS intends to measure specified?</td>
</tr>
<tr>
<td></td>
<td>b. Is (are) the group(s) for which the COS is (are) intended specified?</td>
</tr>
<tr>
<td></td>
<td>c. Is the purpose of the COS specified?</td>
</tr>
<tr>
<td></td>
<td>1.2 Is the theory underlying the COS described?</td>
</tr>
<tr>
<td></td>
<td>1.3 Is the relevance of the COS's content for measuring the construct(s) justified?</td>
</tr>
<tr>
<td></td>
<td>2. Quality of the COS materials</td>
</tr>
<tr>
<td></td>
<td>2.1 Is the COS complete and clear?</td>
</tr>
<tr>
<td></td>
<td>2.2 Are the items in the COS formulated correctly?</td>
</tr>
<tr>
<td></td>
<td>2.3 Is the scoring system devised in such a way that scoring errors can be avoided?</td>
</tr>
<tr>
<td></td>
<td>3. Quality of the rater manual</td>
</tr>
<tr>
<td></td>
<td>3.1 Is a rater manual available?</td>
</tr>
<tr>
<td></td>
<td>3.2 Are the instructions for raters clear and complete?</td>
</tr>
<tr>
<td></td>
<td>3.3 Is information provided on the applications and limitations of the COS?</td>
</tr>
<tr>
<td></td>
<td>3.4 Is a summary of the research findings published in the manual?</td>
</tr>
<tr>
<td></td>
<td>3.5 Is the degree of expertise required by raters to use the COS specified?</td>
</tr>
<tr>
<td></td>
<td>4. Norms</td>
</tr>
<tr>
<td></td>
<td>4.1 Are norms provided?</td>
</tr>
<tr>
<td></td>
<td>4.2 Are the norms up to date?</td>
</tr>
<tr>
<td></td>
<td>Norm-referenced interpretation</td>
</tr>
<tr>
<td></td>
<td>4.3 Are the norm groups large enough and representative?</td>
</tr>
<tr>
<td></td>
<td>Domain-referenced interpretation</td>
</tr>
<tr>
<td></td>
<td>4.4 Is there sufficient agreement between raters?</td>
</tr>
<tr>
<td></td>
<td>4.5 Have the raters been selected and trained appropriately?</td>
</tr>
<tr>
<td></td>
<td>5. Reliability</td>
</tr>
<tr>
<td></td>
<td>5.1 Is information on the reliability of the COS provided?</td>
</tr>
<tr>
<td></td>
<td>5.2 Are the findings of the reliability research sufficient considering the type of decisions that are intended to be made using the COS score?</td>
</tr>
</tbody>
</table>
### Evaluation framework

#### 5.3 What is the quality of the reliability research?

- a. Are the procedures for computing the reliability coefficients correct?
- b. Are the samples for computing the reliability coefficients consistent with the intended use of the COS?
- c. Is the information provided sufficient to make a substantiated judgment of the reliability of the COS?

#### Part B. Evaluation of the validity argument

**The scoring inference**

**Warrant 1:** The scoring rule(s) is/are (statistically) substantiated

- The COS is based on theory, research, and/or standards
- The scoring rule(s) is/are supported by experts
- The scoring rule(s) is/are supported by teachers
- The scoring rule(s) has/have been tested in (pilot) research
- Statistical analyses support the scoring rule(s)
- The psychometric quality of the items is sufficient

**Warrant 2:** Measures were taken to score accurately and consistently

- For each criterion, scoring rules are available for raters
- Raters are trained in the use of the COS
- Raters are expected to meet a certain level of expertise
- Inter-rater reliability is sufficient
- Observation scores are consistent over time

**Warrant 3:** Attention is dedicated to preventing rater bias

- Attention is paid to preventing rater bias during rater training and/or in the COS manual
- Multiple raters are used to compute an observation score
- Statistical analyses show that raters do not rate specific groups of teachers differently from others
- Inter-rater reliability is sufficient

**The generalization inference**

**Warrant:** Opportunities for generalizations are explicitly described in the COS

- The required number of teacher observations is specified and substantiated
- The observation length is specified and substantiated
- The number of raters is specified and substantiated
- The observation moment is specified and substantiated
Appendix A

- The lesson type is specified and substantiated
- A generalizability study, a reliability study, or IRT analyses has shown that the sample of observations are representative of the assessment domain
- Research into the variation of observed lessons supports the generalizations of the observed score
- Confidence intervals are available and are based on a convention

**The extrapolation inference**

*Warrant: The score in the assessment domain is related to the broader target domain*
- The assessment domain covers a great portion of the target domain (face validity)
- The theoretical framework underlying the COS fits within the target domain
- The observed score is related to other measures within the target domain

**The implication inference**

*Warrant 1: The implications for the observable attribute are appropriate (given the theoretical construct)*
*Warrant 2: The implications for the observable attribute are appropriate (given the statistical analyses)*

Note: * data not used in analyses
Evaluation framework
Appendix B: Search syntax in ERIC

Search syntax: (S1 OR S2 OR S3) AND S4 AND S5 AND YR 1990-2014

S1 = TI ( (observation OR evaluation OR inspection OR supervision) N15 (teach* OR class* OR lesson) ) OR AB ( (observation OR evaluation OR inspection OR supervision) N15 (teach* OR class* OR lesson) ) OR SU ( (observation OR evaluation OR inspection OR supervision) N15 (teach* OR class* OR lesson) )

S2 = TI ( (measure OR system OR protocol OR instrument) N15 (teach* OR class* OR lesson) ) OR AB ( (measure OR system OR protocol OR instrument) N15 (teach* OR class* OR lesson) ) OR SU ( (measure OR system OR protocol OR instrument) N15 (teach* OR class* OR lesson) )

S3 = TI ( ("rat* scale" OR "assessment instrument" OR measures) N15 (teach* OR class* OR lesson) ) OR AB ( ("rat* scale" OR "assessment instrument" OR measures) N15 (teach* OR class* OR lesson) ) OR SU ( ("rat* scale" OR "assessment instrument" OR measures) N15 (teach* OR class* OR lesson) )

S4 = TI ( "primary education" OR "early childhood education" OR "elementary education" OR "primary school" OR "elementary school" OR "grade school" OR "infant school" OR "grade-1" OR "K-12" OR "grade 1" OR "grade 2" OR "grade 3" OR "grade 4" OR "grade 5" OR "grade 6" OR "grade 7" OR "grade 8" OR "grades 1-8" OR "primary grades") OR AB ( "primary education" OR "early childhood education" OR "elementary education" OR "primary school" OR "elementary school" OR "grade school" OR "infant school" OR "grade-1" OR "K-12" OR "grade 1" OR "grade 2" OR "grade 3" OR "grade 4" OR "grade 5" OR "grade 6" OR "grade 7" OR "grade 8" OR "grades 1-8" OR "primary grades") OR SU ( "primary education" OR "early childhood education" OR "elementary education" OR "primary school" OR "elementary school" OR "grade school" OR "infant school" OR "grade-1" OR "K-12" OR "grade 1" OR "grade 2" OR "grade 3" OR "grade 4" OR "grade 5" OR "grade 6" OR "grade 7" OR "grade 8" OR "grades 1-8" OR "primary grades")

S5 = TI ( valid* OR relia* OR unrelia* OR replicab* OR repeated OR “internal consistency” OR “item correlation” OR homogen* OR precision OR selection OR reduction OR “item selection” OR test-retest OR (test AND retest) OR stability OR “factor analysis” OR “factor analyses” OR dimension* OR subscale* OR “interscale correlation*” OR error OR “internal consistency” OR cronbach* OR alpha OR correlation* OR kappa* OR “observer variation” OR interrater OR inter-rater OR
intrarater OR intra-rater OR observer variation OR psychometric OR reproducibility OR coefficient OR homogeneity OR interobserver OR inter-observer OR intraobserver OR intra-observer OR repeatable OR “item discriminant” OR “Item response model” OR IRT OR Rasch OR “Differential item functioning” OR DIF ) OR AB ( valid* OR reliability OR unreliability OR replicability OR repeated OR “internal consistency” OR “item correlation” OR homogeneity OR precision OR selection OR reduction OR “item selection” OR test-retest OR (test AND retest) OR stability OR “factor analysis” OR “factor analyses” OR dimension OR subscale OR “interscale correlation” OR error OR “internal consistency” OR Cronbach* OR alpha OR correlation OR kappa OR “observer variation” OR interrater OR inter-rater OR intrarater OR intra-rater OR observer variation OR psychometric OR reproducibility OR coefficient OR homogeneity OR interobserver OR inter-observer OR intraobserver OR intra-observer OR repeatable OR “item discriminant” OR “Item response model” OR IRT OR Rasch OR “Differential item functioning” OR DIF ) OR SU ( validity OR reliability OR unreliability OR replicability OR repeated OR “internal consistency” OR “item correlation” OR homogeneity OR precision OR selection OR reduction OR “item selection” OR test-retest OR (test AND retest) OR stability OR “factor analysis” OR “factor analyses” OR dimension OR subscale OR “interscale correlation” OR error OR “internal consistency” OR Cronbach* OR alpha OR correlation OR kappa OR “observer variation” OR interrater OR inter-rater OR intrarater OR intra-rater OR observer variation OR psychometric OR reproducibility OR coefficient OR homogeneity OR interobserver OR inter-observer OR intraobserver OR intra-observer OR repeatable OR “item discriminant” OR “Item response model” OR IRT OR Rasch OR “Differential item functioning” OR DIF )
Appendix C

Appendix C: Topics in overview document for each COS

Overview document for Each COS

1. General information
   - Name of the COS
   - Abbreviation
   - COS developers
   - Organization / university / publisher
   - Country
   - Language
   - Date of first publication about the COS
   - Date of last publication about the COS
   - Costs of COS use
   - Teachers that can be observed with the COS
   - The teachers in the research into the COS
   - Scoring on paper or online
   - Goal of the COS
   - Availability of a manual

2. Content of the COS
   - Number of versions
   - Constructs that are measured with the COS
   - Theoretical basis of the COS
   - Development of the COS

3. Observed score
   
   Sample of observations
   - Number of lessons that should be observed to compute a score
   - Number of raters per teacher to compute a score
   - Observation moment(s)
   - Type(s) of lessons that should be observed to compute a score
   - Scoring rules
   - Rating scale
   - Information available for standardizing the observation
### Topics in overview document for each COS

<table>
<thead>
<tr>
<th>Rater</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Rater specifications</td>
</tr>
<tr>
<td>- Rater training</td>
</tr>
<tr>
<td>- Rater qualifications</td>
</tr>
<tr>
<td>- Rater calibration</td>
</tr>
<tr>
<td>- How to deal with rater differences</td>
</tr>
<tr>
<td>- How to prevent rater bias</td>
</tr>
</tbody>
</table>

### Computing a score

- Guidelines for computing a score
- How to combine multiple observations
- Availability of norms
- Availability of research into the stability of the observed score
- Is the observed score combined with other measures?

### 4. Use of the score

- Conclusions based on the observation
- Decisions based on the observation
- Actions based on the observation
- To what extent do stakeholders support the conclusions/decisions/actions based on the observation?
- Is the observation/score discussed with the teacher?

### 5. Research into the COS (for each publication)

#### General information

- Title
- Year
- Authors
- Are the authors connected to the development or the sale of the COS?
- Country
- Type of publication
- Journal
- Version of the COS used

#### Research design

- Research questions
Appendix C

- Research design

**Sample**
- Description of the sample
- Sampling method
- Reasoning for the sample size
- Selection of the lessons

**COS use**
- COS use as intended by the COS developers?

**Results**
- Factor analyses
- Scale/item analyses
- IRR analyses
- Correlations with other measures
- Generalizability study, reliability study, or IRT analyses
- Other
### Appendix D: Description and references of included COSs

<table>
<thead>
<tr>
<th>COS</th>
<th>Description</th>
<th>References in review</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Developmentally Appropriate Practice Template (ADAPT)</td>
<td>Developmentally Appropriate Practices is an approach to teaching grounded in research on how young children develop and learn and in what is known about effective early childhood education. The National Association for the Education of Young Children (NAEYC) developed the guidelines for DAP, on which ADAPT is based. ADAPT consists of three scales: curriculum and instruction, interaction and classroom management.</td>
<td>Gottlieb, 1995; Gottlieb, 1997; Gottlieb &amp; Rasher, 1995; Van Horn &amp; Ramey, 2003; Van Horn &amp; Ramey, 2004</td>
</tr>
<tr>
<td>Classroom Atmosphere Instruction Management Student Engagement Instrument (AIMS instrument)</td>
<td>The Classroom AIMS instrument describes exemplary teaching practices on four scales: atmosphere, instruction, management, and student engagement. The practices are based on research into teachers who maintain high levels of student engagement and literacy improvement in students from a wide range of socio-economic backgrounds.</td>
<td>Deault &amp; Savage, 2013; Roehrig, 2003; Roehrig, Bohn, Turner, &amp; Pressley, 2008; Roehrig &amp; Christesen, 2010</td>
</tr>
<tr>
<td>Adult Style Observation Schedule (ASOS)</td>
<td>The ASOS (or checklist leerkrachtstijl in Dutch) is a Belgian COS that measures teachers’ interpersonal style on three scales: stimulation, autonomy support, and emotional support. The criteria are based on research into experience-based education.</td>
<td>De Bilde et al., 2015; Goossens et al., 2009; Laevers et al., 2009; Lento, 2016; Moens, Smits, Van Droogenbroeck, Speybroeck, &amp; Van Damme, 2010; SIBO, n.d.; Thys, 2016; Van Droogenbroeck, Joosten, Imberechts, &amp; Van Damme, 2010; Van Heddegem, Gadeyne, Vandenberghe, Laevers, &amp; Van Damme, 2004; Van Heddegem, Laevers, &amp; Damme, 2004</td>
</tr>
<tr>
<td>COS Description</td>
<td>References in review</td>
<td></td>
</tr>
<tr>
<td>---------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td><strong>Classroom Assessment Scoring System UE (CLASS UE)</strong></td>
<td>Bill &amp; Melinda Gates foundation, 2012; Downer, Stuhlmans, Schweig, Martinez, &amp; Ruzek, 2015; Kane, Kerr, &amp; Pianta, 2014; Pianta, 2011; Pianta &amp; Hamre, 2009b; Pianta, Hamre, &amp; Mintz, 2012; Teachstone, 2015; Teachstone, n.d.; Vitiello, Hadden, &amp; Teachstrone Policy Group, 2014</td>
<td></td>
</tr>
</tbody>
</table>
### Appendix D

<table>
<thead>
<tr>
<th>COS</th>
<th>Description</th>
<th>References in review</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comprehensive Mathematics Instruction Observation Protocol (CMI protocol)</td>
<td>The CMI framework is an instructional framework to guide teachers in implementing constructivist practices in order to strengthen mathematics understanding and achievement. It is based on a variety of literature reviews. The CMI observation protocol measures the impact of the CMI framework on instructional practice using six scales: launch, explore, discuss, classroom climate, mathematical content, and lesson coherence. The number of quality criteria is dependent on the type of lesson.</td>
<td>Hendrickson, Hilton, &amp; Bahr, n.d.; Womack, 2011; Womack &amp; Hilton, n.d.</td>
</tr>
<tr>
<td>Competence Thermometer (or Competentie Thermometer in Dutch) (CTmeter)</td>
<td>The Competence Thermometer (or Competentie Thermometer in Dutch) is a Dutch COS that is part of a teacher assessment system including classroom observation, an interview, document analyses, and an ability test for teachers. The COS can also be used separately. The COS measures four competencies on four scales: interpersonal, pedagogical, didactic, and organizational competence. The criteria are based on literature reviews and the Dutch inspection framework. Observation results are entered in an application and an automatic report is generated, including a benchmark and development tips.</td>
<td>Competentie Thermometer, n.d.; Competentie Thermometer, n.d.-b</td>
</tr>
<tr>
<td>Classroom Strategies Scale-Observer Form (CSS-Form)</td>
<td>The CSS measures the extent to which teachers use strategies that are considered to contribute to student learning. The strategies are based on extensive literature reviews and meta analyses, and include instructional and behavioral management strategies. For the main part of the instrument, the observer rates to which extent (s)he observed the strategy and to which extent (s)he believes the strategy should have been used by the teacher. The highest discrepancy scores indicate the areas for improvement of the teacher.</td>
<td>Reddy &amp; Dudek, 2013a; Reddy &amp; Dudek, 2013b; Reddy &amp; Dudek, 2014; Reddy, Fabiano, &amp; Dudek, 2013; Reddy, Fabiano, Dudek, &amp; Hsu, 2013a; Reddy, Fabiano, Dudek, &amp; Hsu, 2013b; Reddy, Fabiano, Dudek, &amp; Hsu, 2013c</td>
</tr>
</tbody>
</table>
## Description and references of included COSs

<table>
<thead>
<tr>
<th>COS</th>
<th>Description</th>
<th>References in review</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constructivist Teaching Inventory (CTI)</td>
<td>The CTI was developed to measure the nature and extent of constructivist teaching practices in an elementary school classroom. The instrument measures nine key elements of constructivist teaching on four scales: community of learners, teaching strategies, learning activities, and curriculum - assessment. The criteria were deduced from a review of constructivist literature, another COS called ESTEEM (Burry-Stock, 1995), and from a guide on rating instructional conversations (Rueda, Goldenberg, &amp; Gallimore, 1992).</td>
<td>Greer, 1997; Greer, Hudson, &amp; Wiersma, 1999</td>
</tr>
<tr>
<td>Diagnostic Classroom Observation (DCO)</td>
<td>The DCO (or Vermont Classroom Observation Tool, VCOT) measures teaching quality and student learning. It includes a math/science and a literacy version. Both versions include 3 scales: implementation of the lesson, content of the lesson, and culture of the lesson. The DCO was built on four assumptions about the technology of good instruction which are all based on research.</td>
<td>Constantine et al., 2009; Glazerman et al., 2009; Saginor, 2008</td>
</tr>
<tr>
<td>Dialogic Inquiry Tool (DIT)</td>
<td>The DIT (or Argument Rating Tool, ART) was designed to assess the quality of teacher-student interactions in upper elementary language arts classes. It builds on empirical studies showing the importance of collective argumentation and highlighting behaviors that are productive for the development of higher-order thinking during group discussions. DIT includes two scales: community and inquiry. Each criterion is rated on a 6-point scale on which only scores 5 and 6 are an indication of a dialogue. Most items in the DIT are rated for both teacher behavior and for student behavior to also rate the effect of teacher behavior. The DIT was still under development at the time of our review.</td>
<td>Reznitskaya, Glina, &amp; Oyler, 2012; Reznitskaya, Oyler, &amp; Glina, 2015; Reznitskaya, Wilkinson, Oyler, &amp; Glina, 2015</td>
</tr>
</tbody>
</table>
## Appendix D

<table>
<thead>
<tr>
<th>COS</th>
<th>Description</th>
<th>References in review</th>
</tr>
</thead>
<tbody>
<tr>
<td>Developing Language and Literacy Teaching rubrics (DLLT)</td>
<td>The DLLT was developed to assess early literacy teaching and includes separate rubrics for each of the six major instructional activities that form comprehensive literacy instruction: interactive read aloud, shared reading, guided reading, interactive writing, writing workshop, and word study. In addition, it provides a scale for general characteristics of instruction. The rating scale within the DLLT was designed to represent teacher development that proceeds from procedural aspects of instruction to more expert practice. The DLLT was based on a model of literacy processing and literacy acquisition, principles about the nature of expertise development, and previous research.</td>
<td>Hough et al., 2008; Hough et al., 2013; Hung, 2012; Teachscape, n.d.-a; Teachscape, n.d.-b</td>
</tr>
<tr>
<td>Early Childhood Classroom Observation Measure (ECCOM)</td>
<td>The ECCOM was developed to tap into the nature and quality of academic instruction as well as social climate, resources, and other aspects of effective classrooms. The ECCOM measures 17 aspects. For all aspects, it is described how best practices, as defined in the early childhood and educational literature, would look like. It possesses a social constructivist theoretical orientation. For 15 aspects, there are also descriptions of two forms of inappropriate practices: a teacher-controlled classroom that emphasizes the acquisition of basic academic skills, and a child-controlled classroom. For each aspect, all available columns are scored on a 5-point scale, which indicates the amount of time a specific practice was observed.</td>
<td>ECCOM Coding Manual, n.d.; Hauser-Cram, Sirin, &amp; Stipek, 2003; Lerkkanen et al., 2012; Lerkkanen, Kiuru, et al., 2012; Perry, Donohue, &amp; Weinstein, 2007; Stipek, 2004; Stipek &amp; Byler, 2004</td>
</tr>
</tbody>
</table>
### Description and references of included COSs

<table>
<thead>
<tr>
<th>COS</th>
<th>Description</th>
<th>References in review</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Framework for Teaching (FfT)</strong></td>
<td>FfT provides a framework of teaching quality that can function as a shared understanding and common language of the concept. The framework is meant for professional dialogue based on different measures such as artifacts and observation. The FfT is not presented by its developer as a COSs, however observation is one of the measures that is described by the developer to evaluate the FfT. FfT is also widely used as a COS and therefore included in this review. The FfT is grounded in constructivism and measures aspects that have been documented through empirical studies and theoretical research as promoting improved student learning. The FfT includes four scales about planning and preparation, the classroom environment, instruction, and professional responsibilities. Only the first three scales can be observed.</td>
<td>Bill &amp; Melinda Gates foundation, 2012; Bill &amp; Melinda Gates foundation, 2013; Danielson, 2007; Danielson, 2008; Danielson, 2013; Garrett &amp; Steinberg, 2015; Heneman III, Milanowski, Kimball, &amp; Odden, 2006; Ho &amp; Kane, 2013; Kane, Kerr, &amp; Pianta, 2014; Kane, Taylor, Tyler, &amp; Wooten, 2011; Kimball, White, Milanowski, &amp; Borman, 2004; Mihaly, McCaffrey, Staiger, &amp; Lockwood, 2013; Noell, Brownell, Buzick, &amp; Jones, 2014; Sartain, Stoelinga, &amp; Brown, 2011</td>
</tr>
<tr>
<td><strong>International Comparative Analysis of Learning and Teaching (ICALT)</strong></td>
<td>Based on a literature review on effective teaching and learning, the ICALT includes 32 indicators of teacher behavior across 6 scales: safe and stimulating learning climate, efficient classroom management, quality of instruction, teaching learning strategies, stimulating learning environment, adaptation of teaching to diverse student needs. One additional scale, called student engagement includes 3 items that focus on student behavior. The ICALT was originally developed by European inspectorates, but is being developed further by the University of Groningen in The Netherlands.</td>
<td>Rijksuniversiteit Groningen, n.d.; van de Grift, 2007; van de Grift, 2009; van de Grift, 2014; van de Grift &amp; van der Wal, n.d.; van de Grift, van der Wal, &amp; Torenbeek, 2011</td>
</tr>
<tr>
<td><strong>The extent of working with the Interactive Differentiated Direct Instruction Model [Mate van werken met het Interactief Gedifferentieerd Directe Instructiemodel] (IGDI-model)</strong></td>
<td>The IGDI-protocol is a Dutch COS that was developed in evaluation research in a project directed at ‘early reading and dealing with differences’ (BOV-project). The observation instrument measures the instructional quality as defined by the IGDI-model. In the IGDI-model, three aspects are added to direct instruction: interactive instruction and differentiated instruction by the teacher, and cooperative learning by the students. The IGDI-protocol includes 5 scales: preparatory activities, interactive instruction and guided practice, processing, the closing of the lesson, and feedback.</td>
<td>Houtveen, Mijs, Vernooy, Van de Grift, &amp; Koekebacker, 2003; Houtveen &amp; Mijs, 2004</td>
</tr>
</tbody>
</table>
## Appendix D

<table>
<thead>
<tr>
<th>COS</th>
<th>Description</th>
<th>References in review</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instructional Quality Assessment (IQA)</td>
<td>The IQA is an observation instrument aimed at assessing the quality of reading comprehension and math instruction, drawing on two sources of data: classroom observations and teachers’ assignments of student work. The framework is conceptualized around the principles of learning that integrate pedagogical and subject matter knowledge. The IQA includes dimensions concerning accountable talk, expectations, and academic rigor, among other things. The criteria in the two versions of math and reading differ.</td>
<td>Boston, 2008; Boston, 2012a; Boston, 2012b; Boston, 2012c; Boston &amp; Wolf, 2004; Crosson et al., 2006; Junker et al., 2006; Matsumura, Garnier, &amp; Spybrook, 2012; Matsumura et al., 2006; Rosenquist, Wilhelm, &amp; Smith, 2014; Wilhelm, 2014; Wilhelm &amp; Kim, 2015; Wolf, Crosson, &amp; Resnick, 2005</td>
</tr>
<tr>
<td>Marzano Center Teacher Observation Protocol (Marzano Protocol)</td>
<td>The Marzano Teacher Observation Protocol is part of the Marzano teacher evaluation model. This evaluation model is a synthesis of the research into elements traditionally shown to correlate with student academic achievement and includes four domains: classroom strategies and behaviors (the observation instrument), planning and preparing, reflecting on teaching, and collegiality and professionalism. The observation instruments includes 41 key instructional categories organized into 9 Design Questions and grouped into 3 lessons segments involving routine events, addressing content, and enacted on the spot. Not all 41 competencies have to be observed during a single observation, since the content segment includes competencies concerning the interaction with new knowledge, practicing and deepening new knowledge, and generating and testing hypotheses. The Marzano teacher evaluation model can be used in an online web application called iObservation in which teachers can assess themselves, upload personal development plans, and access aligned professional development resources.</td>
<td>Basileo, Toth, &amp; Kennedy, 2015; Learning Sciences International, n.d.; Learning Sciences Marzano Center, 2016; Marzano, 2007; Marzano, 2010; Marzano, 2013; Marzano, 2014; Marzano, Carbaugh, Rutherford, &amp; Toth, 2013; Marzano &amp; Toth, 2013; Marzano, Toth, &amp; Schooling, n.d.; Marzano Research Laboratory, n.d.-a; Marzano Research Laboratory, n.d.-b; Schooling, Toth, &amp; Marzano, n.d.; Toth et al., 2015</td>
</tr>
<tr>
<td>Mathematics Enhancing Classroom Observation Record (MECOR)</td>
<td>MECOR is based on research that indicates which behaviors correlate with student achievement. Since this research was mainly performed outside the UK (in 1998), MECOR was developed to research this relation within the UK. MECOR includes 8 scales: classroom management techniques, classroom behavior, focus/maintain attention to the lesson, review and practice, skills in questioning, enhancement strategies, variety of teaching methods, and positive classroom climate.</td>
<td>Muijs &amp; Reynolds, 2000; Muijs &amp; Reynolds, 2003; Said, 2013; Schaffer, Muijs, Kitson, &amp; Reynolds, n.d.</td>
</tr>
<tr>
<td>COS</td>
<td>Description</td>
<td>References in review</td>
</tr>
<tr>
<td>------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Mathematical Quality of Instruction (MQI)</td>
<td>The MQI measures the nature of the mathematical content available to students during instruction, as expressed in teacher-student, teacher-content, and student-content interactions. Each 5 or 7.5 minute segment of a videotaped lesson containing mathematical instruction for at least half of the time is coded. Some codes are only assigned to the whole lesson. MQI includes four dimensions (richness of the mathematics, working with students and mathematics, errors and imprecision, common core aligned student practices), and additional whole lesson codes about for example student engagement. The MQI was based on literature, the experience of the authors, and watching small segments of video to build the codes used.</td>
<td>Bacher-hicks, Chin, &amp; Hill, n.d.; Bill &amp; Melinda Gates foundation, 2012; Blazar, 2015; Hill, 2014; Hill et al., 2008; Hill, Charalambous, et al., 2012; Hill, Charalambous, &amp; Kraft, 2012; Hill &amp; Chin, n.d.; Hill, Kapitula, &amp; Umland, 2011; Hill, Umland, Litke, &amp; Kapitula, 2012; Kane et al., 2014; Kelcey, Mcginn, Hill, &amp; Charalambous, n.d.; Learning Mathematics for Teaching Project, 2011; Mcginn, Kelcey, &amp; Hill, n.d.; MQI website</td>
</tr>
<tr>
<td>Quality of Classroom Instruction (QCI)</td>
<td>The QCI was developed in the same study as the RCMIS and was developed to examine the impact of the kindergarten Early Learning in Mathematics (ELM) curriculum on teaching behavior. The QCI measures eight aspects of explicit instructional interactions and was designed to complement a frequency-based instrument, the COSTI-M.</td>
<td>Doabler et al., 2015; Doabler, Nelson, et al., 2014; Early Learning in Mathematics Implementation Fidelity Observation, n.d.</td>
</tr>
<tr>
<td>Rapid Assessment of Teacher Effectiveness (RATE)</td>
<td>In order to make observations more informative, cheaper, and faster than current instruments, the RATE instrument was developed. It was developed to be very predictive of student achievement and thus of teachers' future effectiveness. RATE includes 6 items concerning observable teaching behavior. The instrument is still under development.</td>
<td>Gargani &amp; Strong, n.d.; Gargani &amp; Strong, 2014; Gargani &amp; Strong, 2015; Good &amp; Lavigne, 2015; Strong, 2011; The RATE project, 2015</td>
</tr>
<tr>
<td>Ratings of Classroom Management and Instructional Support (RCMIS)</td>
<td>The RCMIS was developed in the same study as the QCI. The RCMIS served as a more global instrument of instructional quality compared to the QCI. Items in the RCMIS target general features of mathematics instructional quality, including classroom management techniques, the delivery of instruction, and the learning environment. The items were based on an analysis of several observation research projects. The RCMIS was designed to complement a frequency-based instrument, the COSTI-M.</td>
<td>Clarke et al., 2014; Doabler &amp; Clarke, et al., 2014; Doabler &amp; Nelson, 2009a; Doabler &amp; Nelson, 2009b; Doabler, Nelson, et al., 2014; Doabler et al., 2015 B</td>
</tr>
<tr>
<td>COS</td>
<td>Description</td>
<td>References in review</td>
</tr>
<tr>
<td>--------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Science and Mathematics Program Improvement Lesson Observation Protocol (SAMPI observation protocol)</td>
<td>The SAMPI observation protocol measures the quality of lessons in order to improve teaching and learning. It includes four scales for the planning/organization of the lesson, the implementation of the lesson, the content of the lesson, and the classroom culture. A fifth optional scale includes indicators about the use of technology to support the lesson. The instrument is based on Michigan state as well as national teaching and learning standards, and includes different scoring rules for science, mathematics, social studies, and language arts.</td>
<td>SAMPI, 2001a; 2001b; 2003a; 2003b</td>
</tr>
<tr>
<td>Science discourse instrument</td>
<td>The Science Discourse Instrument attempts to conceptualize the nature of classroom discourse that would be a reflection of ambitious and rigorous science instruction. Many scales are based on the IQA math version, but the PRACTICE observation instrument focuses on the instruction of science and was developed further based on the science discourse literature. This observation instrument was in an early stage of development during the review.</td>
<td>Osborne et al., 2015; Osborne et al., 2014; PRACTISE project, n.d</td>
</tr>
<tr>
<td>Teacher Advancement Program observational rubric (TAP observational rubric)</td>
<td>TAP is a teacher evaluation system for teacher development and accountability purposes. When TAP is implemented in schools, it has implications for the career paths of teachers, the professional development within the school, and also for teacher compensation. Teachers can have different roles (career, mentor, or master teacher) with different responsibilities, different evaluation criteria and different salaries. Teachers can earn a bonus based on classroom observations with TAP (3-6 times a year) and student growth measures. After each observation conducted by an administrator or teacher leader, a post-conference is conducted aimed at teacher improvement. The TAP evaluation system includes four domains: instruction, the learning environment, designing and planning instruction, and responsibilities. The TAP is based on research: teacher behavior and teaching models research, teaching strategies research, and teacher qualifications research. Furthermore, the Framework for Teaching and other teacher standards were input for this system.</td>
<td>Barnett &amp; Wills, 2015; Barnett, Wills, &amp; Kirby, 2014; Daley &amp; Kim, 2010; Hudson, 2010; Jerald, 2009; Jerald &amp; Van Hook, 2011; Mann, Leutscher, &amp; Reardon, 2013; National Institute for Excellence in Teaching, 2014; National Institute for Excellence in Teaching, 2016; Schacter &amp; Thum, 2004; Solmon, White, Cohen, &amp; Woo, 2007; TAP, n.d.-a; TAP, n.d.-b; U.S. Department of Education, Institute of Education Sciences, &amp; What Works Clearinghouse, 2015</td>
</tr>
<tr>
<td>COS</td>
<td>Description</td>
<td>References in review</td>
</tr>
<tr>
<td>-----------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Virgilio Teacher Behavior Instrument (VTBI)</td>
<td>The VTBI was developed in order to measure the major teacher effectiveness characteristics described in the literature with a manageable number of items in an easy to administer instrument. The items in the instrument measure 10 classroom management, instruction, and classroom climate competencies.</td>
<td>Creemers, Reynolds, &amp; Swint, 1996; Creemers, Stringfield, &amp; Guldenmond, 2002; Crone &amp; Teddlie, 1995; Teddlie, Creemers, Kyriakides, Muijs, &amp; Yu, 2006; Teddlie, Virgilio, &amp; Oescher, 1990; Virgilio, 1987; Virgilio, Teddlie, &amp; Oescher, 1991</td>
</tr>
</tbody>
</table>
## Appendix E: Impact! items

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Items students</th>
<th>Items teachers</th>
<th>Items raters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Learning goals</td>
<td>The teacher clearly indicated what I was going to learn.</td>
<td>I clearly indicated what the students were going to learn.</td>
<td>The teacher clearly indicated what the students were going to learn.</td>
</tr>
<tr>
<td>2. Explanation</td>
<td>The teacher explained the subject matter in such a way that I understood it well.</td>
<td>I explained the subject matter in such a way that the students understood it well.</td>
<td>The teacher explained the subject matter in such a way that the students understood it well.</td>
</tr>
<tr>
<td>3. Prior knowledge</td>
<td>The teacher connected what I already knew with the lesson well.</td>
<td>I connected what students already knew with the lesson well.</td>
<td>The teacher connected what students already knew with the lesson well.</td>
</tr>
<tr>
<td>4. Checking for understanding</td>
<td>The teacher checked whether I understood the subject matter well.</td>
<td>I checked whether the students understood the subject matter well.</td>
<td>The teacher checked whether the students understood the subject matter well.</td>
</tr>
<tr>
<td>5. Asking questions</td>
<td>The teacher asked questions about the subject matter, which made me think.</td>
<td>I asked questions about the subject matter, which made the students think.</td>
<td>The teacher asked questions about the subject matter, which made the students think.</td>
</tr>
<tr>
<td>6. Additional explanation</td>
<td>If I did not understand the subject matter, the teacher made sure I understood it.</td>
<td>If students did not understand the subject matter, I made sure they understood it.</td>
<td>If students did not understand the subject matter, the teacher made sure they understood it.</td>
</tr>
<tr>
<td>7. Classroom work climate</td>
<td>The teacher made sure we worked hard during the lesson.</td>
<td>I made sure the students worked hard during the lesson.</td>
<td>The teacher made sure the students worked hard during the lesson.</td>
</tr>
<tr>
<td>8. Safe climate</td>
<td>The teacher made sure I was not afraid to say if I did not understand something.</td>
<td>I made sure that students were not afraid to say if they did not understand something.</td>
<td>The teacher made sure that students were not afraid to say if they did not understand something.</td>
</tr>
<tr>
<td>9. Pleasant climate</td>
<td>The teacher created a good classroom climate.</td>
<td>I created a good classroom climate.</td>
<td>The teacher created a good classroom climate.</td>
</tr>
<tr>
<td></td>
<td>Impact! items</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>--------------</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>10</td>
<td>Lesson pace</td>
<td>The pace of the lesson was good for me.</td>
<td>The pace of the lessons was good for all students.</td>
</tr>
<tr>
<td>11</td>
<td>Feedback</td>
<td>If I gave an incorrect answer, the teacher explained why it was wrong.</td>
<td>If students gave an incorrect answer, I explained why it was wrong.</td>
</tr>
<tr>
<td>12</td>
<td>Active participation</td>
<td>The teacher ensured my active participation in the lesson.</td>
<td>I ensured active participation of the students in the lesson.</td>
</tr>
<tr>
<td>13</td>
<td>Time for assignments</td>
<td>The teacher gave enough time to work on the assignments in the lesson.</td>
<td>I gave enough time to work on the assignments in the lesson.</td>
</tr>
<tr>
<td>14</td>
<td>Summarizing lesson content</td>
<td>At the end of the lesson, the teacher summarized what we had learned.</td>
<td>At the end of the lesson, I summarized what students had learned.</td>
</tr>
<tr>
<td>15</td>
<td>Ability to complete assignments</td>
<td>I now can complete assignments about the subject matter on my own.</td>
<td>Students can now complete assignments about the subject matter on their own.</td>
</tr>
<tr>
<td>16</td>
<td></td>
<td>The teacher’s lessons will become even better if he/she ...</td>
<td></td>
</tr>
</tbody>
</table>

Note: * students had the opportunity to answer in the response category ‘not applicable’.
Appendix F: Bugs code

# N teacher
# KK items
# MM categories
# NN students
# K raters
# M Lessons
# NDIM three dimensions
model {
  # model students
  for (i in 1 : N) {
    for (s in 1 : NN) {
      for (t in 1 : M) {
        for (k in 1 : KK) {
          for (j in 1 : MM) {
            num01[i,s,t,k,j] <- exp((j-1)*alpha[k]*theta_ist[i,s,t]-beta[k,j])
            den01[i,s,t,k] <- sum(num01[i,s,t,k,j])
          }
        }
      }
    }
  }
  for (i in 1 : N) {
    for (s in 1 : NN) {
      for (t in 1 : M) {
        for (k in 1 : KK) {
          for (j in 1 : MM) {
            prob0[i,s,t,k,j] <- num01[i,s,t,k,j] / den01[i,s,t,k]
          }
        }
      }
    }
  }
  XLL[i,t,s,k] ~ dcat(prob0[i,s,t,k,1:MM])
  exp.theta_ist[i,s,t] <- theta[i,1] + gamma.is[(i-1)*NN+s,t] + v.t[t] + v.it[i,t]
  theta_ist[i,s,t] ~ dnorm(exp.theta_ist[i,s,t], tau.ist)
  gamma.is[(i-1)*NN+s,t] ~ dnorm( v.is[(i-1)*NN+s], tau.vis)
}

# model external raters
for (i in 1 : N) {
  for (t in 1 : M) {
    for (r in 1 : K) {
      for (k in 1 : KK) {
        for (j in 1 : MM) {
          num1[i,t,r,k,j] <- exp((j-1)*alpha[items[k]]*theta_itr[i,t,r]-beta[items[k],j])
          den1[i,t,r,k] <- sum(num1[i,t,r,k,j])
        }
      }
    }
  }
  XLL[i,t,s,k] ~ dcat(prob0[i,s,t,k,1:MM])
  exp.theta_ist[i,s,t] <- theta[i,1] + gamma.is[(i-1)*NN+s,t] + v.t[t] + v.it[i,t]
  theta_ist[i,s,t] ~ dnorm(exp.theta_ist[i,s,t], tau.ist)
  gamma.is[(i-1)*NN+s,t] ~ dnorm( v.is[(i-1)*NN+s], tau.vis)
}
for (t in 1 : M) {
  for (r in 1 : K) {
    for (k in 1 : KK) {
      for (j in 1 : MM) {
        prob[i,t,r,k,j] <- num1[i,t,r,k,j] / den1[i,t,r,k]  
        XOBS[i,t,r,k] ~ dcat(prob[i,t,r,k,1:MM])   }
    exp.theta_itr[i,t,r] <- theta[i,2]+u.t[t]+u.r[r]+u.it[i,t]+u.ir[i,r]+u.tr[t,r]
    theta_itr[i,t,r] ~ dnorm(exp.theta_itr[i,t,r], tau.itr )   }}

  # model teachers  
  for (i in 1 : N) {
    for (t in 1 : M) {
      for (k in 1 : KK) {
        for (j in 1 : MM) {
          num2[i,t,k,j] <- exp((j-1)*alpha[k]*theta_it[i,t]-beta[k,j])    
        den2[i,t,k] <- sum(num2[i,t,k, ])  
        prob2[i,t,k,j] <- num2[i,t,k,j] / den2[i,t,k]  
        XLK[i,t,k] ~ dcat(prob2[i,t,k,1:MM])   }
    exp.theta_it[i,t] <- theta[i,3]    
    theta_it[i,t] ~ dnorm(exp.theta_it[i,t], tau.i )   
}}

  # model covariance  

  # Priors covariance 
  R[1:NDIM,1:NDIM] ~ dwish(Omega[1:NDIM,1:NDIM], NDIM) 
  IR[1:NDIM,1:NDIM] <- inverse(R[1:NDIM,1:NDIM])
  for (j in 1 : NDIM ) { mu[j] ~ dnorm(0,1) }
  for (j in 1 : NDIM ) { Omega[j,j]<- 1 }
  for ( k in 1:j-1 ) {Omega[j,k] <- 0}
  for ( k in j+1:NDIM ) {Omega[j,k] <- 0}     

  # Priors IRT model  
  for (k in 1:KK) {  
    alog[k]~dnorm(0,25)
Appendix F

alpha[k]<-exp(alog[k])
beta[k,1] <- 0
for (j in 2:MM) { beta[k,j] ~ dnorm(0, 1)
location[k,j] <- (beta[k,j]-beta[k,j-1]) / alpha[k] }}

# Priors GT model
for (t in 1:M)  { v.t[t] ~ dnorm(0, tau.vt) }
for (i in 1 : N) { for (t in 1 : M) { v.it[i,t] ~ dnorm(0, tau.vit)}
for (s in 1:NN) { v.is[(i-1)*NN+s] ~ dnorm(0,1) }}
for (t in 1:M) { u.t[t] ~ dnorm(0, tau.t) }
for (r in 1:K) { u.r[r] ~ dnorm(0, tau.r) }
for (i in 1 : N) { for (t in 1:M) { u.it[i,t] ~ dnorm(0, tau.it) }}
for (i in 1 : N) { for (r in 1:K) { u.ir[i,r] ~ dnorm(0, tau.ir) }}
for (t in 1 : M) { for (r in 1:K) { u.tr[t,r] ~ dnorm(0, tau.tr) }}

tau.itr ~ dgamma(25,5)
var.itr <- 1/tau.itr
tau.ist ~ dgamma(25,5)
var.ist <- 1/tau.ist
tau.t ~ dgamma(25,5)
var.t <- 1/tau.t
tau.r ~ dgamma(25,5)
var.r <- 1/tau.r
tau.vt ~ dgamma(25,5)
var.vt <- 1/tau.vt
tau.vis ~ dgamma(25,5)
var.vis <- 1/tau.vis
tau.vit ~ dgamma(25,5)
var.vit <- 1/tau.vit
tau.it ~ dgamma(25,5)
var.it <- 1/tau.it
tau.ir ~ dgamma(25,5)
var.ir <- 1/tau.ir
tau.tr ~ dgamma(25,5)
var.tr <- 1/tau.tr
tau.i ~ dgamma(25,5)
var.i<- 1/tau.i
REL_LL   <- IR[1,1] / (IR[1,1] + (var.vt/3) + (var.vit/3) + (var.vis/20) +(var.ist/60))
AGREE_OBS<- IR[2,2] / (IR[2,2]+var.r/K+var.t/M+var.it/M+var.ir/K+var.tr/(M*K)+var.itr/(M*K))
REL_OBS <- IR[2,2] / (IR[2,2]+var.it/M+var.ir/K+var.itr/(M*K))
}
Appendix G

Appendix G: Item parameter estimates and information values for three groups of raters

The table displayed below gives the estimates of the item parameters and the information value of the items. The estimates are obtained in separate analyses for the three groups of raters. In each of the three analyses, the average of the latent proficiency variable was set equal to zero to identify each model. Therefore, the absolute values of the estimates are not on a common scale yet and only their correlations are informative. They are provided in the main text.

The item location parameters $\overline{\delta}_k$ are the means of the location parameters of individual item category parameters, that is

$$\overline{\delta}_k = \frac{1}{M} \sum_{m=1}^{M} \delta_{km}$$

These parameters can be seen as item difficulty parameters, that is, they gauge the extent to which an item is endorsed. For instance, for the external raters, the item about the quality of provided explanations has a high average observed score and the item about summarizing the lesson content has a low average observed score. Together, the item parameters determine the information value of an item. The information value is a local measure that signifies the contribution of the item to the reliability at a certain point of the latent scale. The higher the information value, the higher the contribution. The values given in the table are computed at the mean of the distribution of each group of raters.

The results show that for the external raters, the item about active participation has the highest contribution. The items about learning goals and summarizing the lesson content contribute very little. For the students, the items about the quality of provided explanations, checking for understanding, additional explanation, pleasant climate, and active participation function very well, that is, they contribute much to the reliability. The items regarding time for assignments and summarizing the lesson content have the lowest contribution. For the teachers, the item about learning goals and time for assignments contribute little, and the item about classroom work climate contributes most.
Table G.1 Item parameter estimates and information values for three groups of raters

<table>
<thead>
<tr>
<th>Item</th>
<th>External Raters</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha_k$</td>
<td>$\delta_k$</td>
<td>Info</td>
<td>$\alpha_k$</td>
<td>$\delta_k$</td>
<td>Info</td>
<td>$\alpha_k$</td>
<td>$\delta_k$</td>
</tr>
<tr>
<td>Learning goals</td>
<td>0.54</td>
<td>-0.32</td>
<td>0.14</td>
<td>1.46</td>
<td>-4.00</td>
<td>0.68</td>
<td>0.35</td>
<td>-1.29</td>
</tr>
<tr>
<td>Explanation</td>
<td>2.89</td>
<td>-14.78</td>
<td>0.55</td>
<td>2.10</td>
<td>-5.20</td>
<td>1.13</td>
<td>2.46</td>
<td>-3.49</td>
</tr>
<tr>
<td>Prior knowledge</td>
<td>1.65</td>
<td>-3.13</td>
<td>0.41</td>
<td>1.63</td>
<td>-3.84</td>
<td>0.71</td>
<td>1.59</td>
<td>-3.90</td>
</tr>
<tr>
<td>Checking for understanding</td>
<td>1.45</td>
<td>-3.36</td>
<td>0.67</td>
<td>1.83</td>
<td>-3.49</td>
<td>1.05</td>
<td>1.19</td>
<td>-2.78</td>
</tr>
<tr>
<td>Asking questions</td>
<td>1.66</td>
<td>-2.91</td>
<td>0.85</td>
<td>1.26</td>
<td>-2.51</td>
<td>0.69</td>
<td>0.63</td>
<td>-0.12</td>
</tr>
<tr>
<td>Additional explanation</td>
<td>1.41</td>
<td>-4.61</td>
<td>0.39</td>
<td>2.06</td>
<td>-5.00</td>
<td>1.09</td>
<td>1.11</td>
<td>-2.45</td>
</tr>
<tr>
<td>Classroom work climate</td>
<td>1.42</td>
<td>-3.66</td>
<td>0.72</td>
<td>1.77</td>
<td>-4.22</td>
<td>0.90</td>
<td>1.95</td>
<td>-0.39</td>
</tr>
<tr>
<td>Safe climate</td>
<td>1.66</td>
<td>-5.90</td>
<td>0.28</td>
<td>1.80</td>
<td>-4.93</td>
<td>0.83</td>
<td>1.09</td>
<td>-2.80</td>
</tr>
<tr>
<td>Pleasant climate</td>
<td>1.26</td>
<td>-4.79</td>
<td>0.29</td>
<td>2.09</td>
<td>-5.01</td>
<td>1.02</td>
<td>2.07</td>
<td>-4.07</td>
</tr>
<tr>
<td>Lesson pace</td>
<td>1.30</td>
<td>-4.73</td>
<td>0.26</td>
<td>1.70</td>
<td>-4.29</td>
<td>0.82</td>
<td>1.27</td>
<td>-1.41</td>
</tr>
<tr>
<td>Feedback</td>
<td>1.32</td>
<td>-3.51</td>
<td>0.31</td>
<td>1.49</td>
<td>-3.77</td>
<td>0.67</td>
<td>1.79</td>
<td>2.66</td>
</tr>
<tr>
<td>Active participation</td>
<td>2.33</td>
<td>-5.03</td>
<td>1.13</td>
<td>2.05</td>
<td>-4.56</td>
<td>1.05</td>
<td>2.32</td>
<td>-1.61</td>
</tr>
<tr>
<td>Time for assignments</td>
<td>0.79</td>
<td>-2.74</td>
<td>0.20</td>
<td>1.11</td>
<td>-2.97</td>
<td>0.51</td>
<td>0.31</td>
<td>-3.25</td>
</tr>
<tr>
<td>Summarizing lesson content</td>
<td>0.38</td>
<td>2.88</td>
<td>0.08</td>
<td>0.91</td>
<td>-1.15</td>
<td>0.50</td>
<td>0.52</td>
<td>0.09</td>
</tr>
<tr>
<td>Ability to complete assignments</td>
<td>2.57</td>
<td>-0.87</td>
<td>0.29</td>
<td>1.67</td>
<td>-4.60</td>
<td>0.82</td>
<td>1.99</td>
<td>-5.44</td>
</tr>
</tbody>
</table>
Summary in Dutch
Nederlandse samenvatting
Nederlandse samenvatting
**Introductie**

De kwaliteit van de leraar is van groot belang voor het leren van leerlingen. Het meten van deze kwaliteit is daarmee ook belangrijk, zodat er alleen leraren voor de klas komen van voldoende niveau, de leraren die voor de klas staan weten op welke punten zij nog kunnen verbeteren, en tijdig HRM beslissingen genomen kunnen worden als de kwaliteit van de leraar beneden een gewenst niveau is. Het meten van lerarenkwaliteit kan op diverse manieren. Te denken valt aan het gebruiken van leerlingresultaten, het meten van leerlingpercepties of het doen van lesobservaties. Elke methode heeft voor- en nadelen.

Dit proefschrift focust zich voornamelijk op lesobservaties en de kwaliteit en kenmerken van de instrumenten die daarbij gebruikt worden. In de definitie van lesobservatie-instrumenten die we daarbij aanhouden, bestaat een lesobservatie-instrument uit drie aspecten:

1. de scoringsinstrumenten met de gedragingen of items die gescoord worden;
2. de maatregelen die getroffen worden met betrekking tot de betrouwbaarheid van de observatoren, zoals de training van observatoren of de beschikbaarheid van een handleiding;
3. de specificaties met betrekking tot de steekproef, zoals de kenmerken van de lessen die geobserveerd moeten worden en het aantal lessen dat geobserveerd moet worden om de gewenste uitspraken te kunnen doen over de kwaliteit van een leraar.

Recent onderzoek laat zien dat het verkrijgen van betrouwbare en valide scores met een lesobservatie-instrument niet vanzelfsprekend is. Aandacht voor alle drie de aspecten van een lesobservatie-instrument zijn daarvoor van groot belang. In de onderwijspraktijk en het onderwijsonderzoek bleek lang niet altijd aandacht voor deze drie aspecten, en bleek er een grote kloof te bestaan tussen wat we weten over lesobservaties vanuit de theorie en hoe dit wordt ingezet in de praktijk.

In de eerste studie in dit proefschrift zijn we in de literatuur op zoek gegaan naar welke kwaliteitsaspecten er voor lesobservatie-instrumenten genoemd worden.
Aan welke kwaliteitskenmerken moet een lesobservatie-instrument voldoen?

We hebben gezocht naar kwaliteitscriteria voor lesobservatie-instrumenten in de literatuur over lesobservatie(instrumenten), algemene literatuur over tests en assessments, en literatuur over validiteit. Dit resulteerde in een evaluatiekader dat bestond uit drie delen, zoals te zien is in onderstaand figuur.

<table>
<thead>
<tr>
<th>BELANGRIJKE KENMERKEN</th>
<th>EVALUATIEKADER</th>
<th>VALIDITEIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scorings-instrumenten</td>
<td>Beschikbaarheid onderzoek naar betrouwbaarheid</td>
<td>Voldoende bewijs voor validiteit van...</td>
</tr>
<tr>
<td>Omschrijving doel</td>
<td>Evaluatie van beschikbare gegevens</td>
<td>... scoren</td>
</tr>
<tr>
<td>Theoretische basis</td>
<td>Evaluatie van kwaliteit van onderzoek</td>
<td>... generaliseren</td>
</tr>
<tr>
<td>Kwaliteits items</td>
<td></td>
<td>... extrapoleren</td>
</tr>
<tr>
<td>Maatregelen voor</td>
<td></td>
<td>... impliceren</td>
</tr>
<tr>
<td>betrouwbaarheid</td>
<td></td>
<td></td>
</tr>
<tr>
<td>observatoren</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Richtlijnen voor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gebruik</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figuur 1. Overzicht evaluatiekader

Het eerste deel van het evaluatiekader gaat in op belangrijke kenmerken van lesobservatie-instrumenten. Bijvoorbeeld de motivatie voor de gedragingen of items die gescoord worden met het instrument (de theoretische basis). Dit kunnen gedragingen zijn die in onderzoek samenhangen met betere leerresultaten. Het tweede gedeelte van het kader gaat in op de gegevens die beschikbaar zijn over de betrouwbaarheid van de scores die verkregen kunnen worden met het lesobservatie-instrument. Dit geeft een indicatie van de contexten waarin het instrument tot betrouwbare scores heeft geleid. Het is hierbij ook van belang om de kwaliteit van het onderzoek naar de betrouwbaarheid mee te nemen.

Met het derde gedeelte van het kader wordt aan de hand van de argumentatieve benadering van validiteit gekeken naar de validiteit van de scores, zoals weergegeven in Figuur 2.
Figuur 2. Argumentatieve benadering van validiteit

Helemaal links in Figuur 2 staan de observaties die gedaan worden (de steekproef) en helemaal rechts staan de conclusies die getrokken worden of beslissingen die genomen worden op basis van deze observaties. De blokken ertussen representeren de redeneerlijn die over het algemeen wordt gevolgd om van de observaties tot de conclusies/beslissingen te komen. Daarin worden vier stappen onderscheiden (scoren, generaliseren, extrapoleren en impliceren) en met het evaluatiekader wordt het bewijs voor validiteit van elk van deze vier stappen geëvalueerd.

1. Scoren
Allereerst wordt bekeken of er voldoende bewijs is voor het scoren met het instrument, dat wil zeggen of het valide is om op basis van de observatie(s) tot een score te komen. Te denken valt aan bewijs voor de items in het instrument, bijvoorbeeld als deze zijn voorgelegd aan experts, of aan bewijs voor de betrouwbaarheid van de observatoren, bijvoorbeeld als hiervoor training beschikbaar is.

2. Generaliseren
Als tweede wordt bekeken of er voldoende bewijs is voor het generaliseren van de geobserveerde score, als dit aan de orde is. Vaak is het niet het doel om uitspraken te doen over alleen de geobserveerde lessen, maar over de kwaliteit in het algemeen (score in het assessmentdomein). Met het evaluatiekader kijken we hier naar de richtlijnen die hiervoor gegeven worden in het lesobservatie-instrument en hoe deze richtlijnen zijn onderbouwd. Onderzoek kan bijvoorbeeld uitwijzen hoeveel lessen door hoeveel observatoren geobserveerd moeten worden om met het instrument betrouwbaar iets te zeggen over een bredere periode.
3. Extrapoleren
Als derde wordt bekeken of er voldoende bewijs is voor het extrapoleren van de score. Dit is alleen relevant als er op basis van de observatie uitspraken worden gedaan over een breder domein (het praktijk domein) dan dat daadwerkelijk is geobserveerd. Een voorbeeld van extrapoleren is als er alleen geobserveerd is bij rekenlessen, maar er uitspraken gedaan worden over de kwaliteit van een leraar in het algemeen. Relevant bewijs is in dat geval bewijs voor de samenhang tussen rekenlessen en andere lessen die de leraar geeft.

4. Impliceren
Tot slot wordt er gekeken naar het bewijs voor de uitspraken die gedaan worden op basis van de observatie-scores en of bijvoorbeeld beslissingen passend zijn.

Het evaluatiekader is relevant voor instrumentontwikkelaars, omdat het aangeeft welke aspecten van belang zijn bij de ontwikkeling van een lesobservatie-instrument. Het is ook bedoeld voor (potentiële) gebruikers van bestaande instrumenten, om de kwaliteit van deze instrumenten te evalueren en het gebruik van een instrument in de eigen context te evalueren. Zelf hebben we het evaluatiekader gebruikt in de tweede studie om bestaande lesobservatie-instrumenten voor het primair onderwijs te beoordelen.

Wat is de kwaliteit van bestaande lesobservatie-instrumenten voor het primair onderwijs?

Diverse lesobservatie-instrumenten zijn ontwikkeld door onderzoekers, mensen in de praktijk, overheden en commerciële partijen. In de tweede studie zijn we op zoek gegaan naar Nederlandse en Engelse instrumenten die ontwikkeld zijn voor het meten van (een aspect van) lerarenkwaliteit in het primair onderwijs, om deze instrumenten vervolgens te beoordelen met het evaluatiekader. Een voorwaarde was dat er enig onderzoek naar het instrument beschikbaar moest zijn, omdat anders de onderdelen betrouwbaarheid en validiteit niet beoordeeld konden worden. Opvallend was dat de meeste lesobservatie-instrumenten, beschikbaar voor het primair onderwijs in Nederland, hierdoor afvielen. Deze instrumenten werden ten tijde van onze zoektocht wel aangeboden/verkocht aan scholen, maar er was geen enkel onderzoek dat liet zien dat de instrumenten betrouwbare/valide scores op zouden kunnen leveren.
Nederlandse samenvatting

In totaal voldeden 27 lesobservatie-instrumenten aan onze zoekcriteria. Elk instrument is beoordeeld door twee getrainde beoordelaars met het evaluatiekader. De resultaten laten zien dat de beoordelaars van de drie aspecten in onze definitie van lesobservatie-instrumenten, het meest positief waren over de scoringsinstrumenten. In de meeste instrumenten werd ook aandacht besteed aan de betrouwbaarheid van de observatoren (het tweede aspect in de definitie). Echter, bij de meeste instrumenten betwijfelde ten minste één beoordelaar of deze maatregelen er voldoende voor zouden zorgen dat observatoren accuraat en betrouwbaar zouden kunnen scoren met het instrument. Het minst positief waren de beoordelaars over het derde aspect in de definitie: instrument-ontwikkelaars waren volgens hen weinig expliciet over de specificaties met betrekking tot de steekproef. Voor de meeste lesobservatie-instrumenten gold ten slotte ook dat de beoordelaars vaak ook niet overtuigd waren van het bewijs voor betrouwbaarheid en validiteit van de scores. Dit kwam vooral door het ontbreken van gedegen onderzoek.

Deze resultaten geven geen positief beeld van de kwaliteit van de lesobservatie-instrumenten die bij de start van de review beschikbaar waren. Instrumentontwikkelaars lijken vooral te focussen op het ontwikkelen van de scoringsinstrumenten. Mogelijk zijn er onvoldoende middelen als geld en tijd om ook aandacht te besteden aan de twee andere aspecten in onze definitie en gedegen onderzoek uit te voeren naar het instrument. Een andere mogelijke verklaring is dat instrumentontwikkelaars en gebruikers van de instrumenten zich onvoldoende bewust zijn van het belang hiervan.

De eerste twee studies in dit proefschrift laten zien dat het verkrijgen van valide observatie-scores geen gemakkelijke taak is. In de derde studie hebben we daarom onderzocht in hoeverre observatie scores samenhangen met twee andere metingen van lerarenkwaliteit.

Beoordelen externe observatoren, leraren en leerlingen leraarkwaliteit hetzelfde?

Er is nog veel onduidelijkheid over hoe de resultaten van verschillende methoden om lerarenkwaliteit te meten met elkaar samenhangen. In deze studie hebben we onderzocht hoe de beoordelingen van een les door getrainde observatoren samenhangen met de percepties van leerlingen en leraren over dezelfde les. Hiervoor zijn data verzameld bij 25 wiskundeleraren verspreid over Nederland en hun 608 havo 3 leerlingen. Bij alle leraren zijn data verzameld over drie lessen.
Nederlandse samenvatting

De perceptie van de leerlingen over de leskwaliteit is gemeten met de Impact! tool. Dit is een applicatie waarmee leerlingen aan het einde van een les de leraar feedback kunnen geven over de gegeven les. De leerlingen beantwoorden in deze applicatie 15 gesloten vragen op een 4-puntsschaal. Deze vragen zijn gebaseerd op wetenschappelijke literatuur en gaan over kenmerken van een effectieve les, bijvoorbeeld of de uitleg begrijpelijk was voor de leerlingen. In dit onderzoek hebben de leraren dezelfde 15 vragen beantwoord aan het einde van de les. Alle 75 lessen zijn ook opgenomen op video en achteraf beoordeeld door drie getrainde observatoren met dezelfde 15 items.

De leraren waren in dit onderzoek gemiddeld het meest positief over de leskwaliteit, gevolgd door de leerlingen. De externe observatoren waren gemiddeld een stuk minder positief. De beoordelingen van de leerlingen en de leraren hingen echter helemaal niet met elkaar samen, zoals te zien is in Figuur 3. Hetzelfde geldt voor de beoordelingen van de externe observatoren en die van de leraren. We vonden een matige samenhang tussen de beoordelingen van de externe observatoren en die van de leerlingen.

Figuur 3. Overeenstemming observatoren, leerlingen en leraren
Nederlandse samenvatting

Mogelijke verklaringen voor deze resultaten zijn dat:

- observatoren, leerlingen en leraren een andere visie hebben gehad op bepaalde aspecten, zoals wat het benoemen van een lesdoel moet inhouden voor een voldoende score of wanneer er hard genoeg gewerkt is tijdens een les voor een voldoende score;
- de scores beïnvloed zijn door andere factoren, bijvoorbeeld wanneer leraren zichzelf hoger beoordeeld hebben, omdat ze wisten dat onderzoekers naar de antwoorden zouden kijken (sociale wenselijkheid) of wanneer leerlingen de les anders beoordeeld hebben door de invloed van andere factoren, zoals cijfers die ze net gekregen hebben;
- het moeilijk is voor externe observatoren en leraren om het effect van lesaspecten op leerlingen te observeren. Bij de 15 vragen die de beoordelaars hebben ingevuld zaten diverse vragen die gingen over de ervaringen van leerlingen, bijvoorbeeld of ze zich veilig voelden in de klas en of de leraar vragen heeft gesteld die hen aan het denken hebben gezet. Het kan zijn dat dergelijke vragen moeilijk te observeren zijn en het beste aan leerlingen zelf gesteld kunnen worden.

De resultaten laten zien dat het bij het meten van leskwaliteit belangrijk is om na te denken welke meetmethode het meest geschikt is, omdat elke methode kan leiden tot een ander resultaat en mogelijk niet alle aspecten van leskwaliteit met dezelfde methode het beste gemeten kunnen worden.

Aanbevelingen op basis van het onderzoek

Alle drie de studies laten zien dat het meten van leskwaliteit door middel van observatie complex is. Op basis van de resultaten uit de studies kunnen een aantal aanbevelingen gedaan worden voor instrumentontwikkelaars, gebruikers van lesobservatie-instrumenten en de overheid.

Ontwikkelaars lesobservatie-instrumenten

De review laat zien dat de meeste lesobservatie-instrumenten die ten tijde van de review werden gebruikt in de praktijk en in onderzoek, niet aan de standaarden voldeden die in het evaluatiekader omschreven worden. Om ervoor te zorgen dat lesobservaties valide scores opleveren, en de tijdsinvestering die gedaan wordt om dergelijke scores te verkrijgen zinvol is, is het van belang dat instrument-ontwikkelaars meer aandacht krijgen voor:
- alle drie de componenten in onze definitie van lesobservatie-instrumenten;
Nederlandse samenvatting

- onderzoek naar betrouwbaarheid en validiteit van de verkregen scores met de instrumenten;
- het omschrijven van duidelijke richtlijnen voor het gebruik van de instrumenten op basis van dit onderzoek.

Instrumentontwikkelaars kunnen op die manier de schakel vormen tussen wetenschappelijk onderzoek en hoe losobservatie-instrumenten worden gebruikt in de praktijk. Hiervoor is het uiteraard wel nodig dat zij zich bewust zijn van het belang van de kwaliteitsstandaarden in het evaluatiekader en van de complexiteit van het ontwikkelen van een losobservatie-instrument.

Gebruikers losobservatie-instrumenten
Het is van belang dat ook de gebruikers van losobservatie-instrumenten op de hoogte zijn van de kwaliteitsstandaarden in het evaluatiekader en het belang hiervan inzien. Hierdoor kiezen zij hopelijk alleen nog instrumenten die hieraan voldoen en houden zij zich aan de richtlijnen die door instrumentontwikkelaars worden opgesteld voor het gebruik ervan op basis van onderzoek. Dit kan het verkrijgen van observatiescores duurder en tijdsintensiever maken, bijvoorbeeld omdat observatoren getraind moeten worden of omdat meerdere observaties gedaan moeten worden. Het is echter van groot belang dat gebruikers van losobservatie-instrumenten inzien dat het verkrijgen van onbetrouwbare of niet valide scores geen meerwaarde heeft en een verspilling is van de middelen die ervoor nodig zijn om dergelijke scores te verkrijgen.

Hoewel instrumentontwikkelaars in onze ogen de eerste verantwoordelijkheid hebben om onderzoek te doen naar de validiteit en betrouwbaarheid van de observatiescores met een bepaald instrument, is het ook van belang dat gebruikers van deze instrumenten goed kijken naar het instrumentgebruik in hun eigen context. Deze context kan de betrouwbaarheid en validiteit van de scores beïnvloeden, bijvoorbeeld als leraren door observatoren worden geobserveerd die voor hen bekend zijn, zoals de schoolleider. In het derde deel van het evaluatiekader staan vragen die gebruikers daarbij kunnen stellen.

Op basis van de resultaten in de derde studie kan het voor gebruikers van losobservatie-instrumenten ook relevant zijn om losobservaties met andere meetmethoden te combineren.
Overheid

Het is sterk de vraag of meer bewustzijn van het belang van de kwaliteitsstandaarden en de complexiteit ervan bij ontwikkelaars en gebruikers van lesobservatie-instrumenten voldoende is voor de benodigde veranderingen in het gebruik van lesobservaties.

Er zijn meerdere manieren waarop de overheid het gebruik van lesobservatie-instrumenten die aan de kwaliteitsstandaarden voldoen, kan stimuleren, namelijk door:

1. het ontwikkelen van een beoordelingssysteem voor lesobservatie-instrumenten. De PO-Raad heeft een dergelijk beoordelingssysteem ontwikkeld voor observatie-instrumenten in het primair onderwijs, waarbij het evaluatiekader uit de eerste studie ook als input gediend heeft;
2. het mogelijk maken van onafhankelijke beoordelingen van de lesobservatie-instrumenten, zoals de COTAN doet voor de kwaliteit van tests;
3. het ontwikkelen van een database met alleen de instrumenten die aan de kwaliteitsstandaarden voldoen.

Als duidelijk is welke instrumenten aan de kwaliteitsstandaarden uit het evaluatiekader voldoen, kan de overheid van scholen verlangen om alleen middelen te besteden aan het gebruik van dergelijke instrumenten.
Acknowledgements
Dankwoord
Dankwoord
Dankwoord

De ambitie om een proefschrift te schrijven, kreeg ik toen ik na mijn studie Onderwijskunde aan het werk ging bij de directie Kennis van de Onderwijsinspectie. Ik werd hierin gesteund door mijn leidinggevenden en uiteindelijk kreeg ik de mogelijkheid om te starten met dit proefschrift in de academische werkplaats Kwaliteit in de klas. Ik wil de inspectie bedanken voor deze kans en wil daarbij in het bijzonder Bert, Marian, Rinske, Frans en Dré bedanken voor hun vertrouwen in mij en hun inzet hiervoor.

Het uitvoeren van dit promotieonderzoek en het schrijven van het proefschrift was niet gelukt zonder de steun van anderen en was zeker ook niet zo leuk en leerzaam geweest zonder de samenwerking met anderen.

Allereerst wil ik mijn promotor Adrie bedanken, want zonder hem was dit proefschrift er niet geweest. Adrie, bedankt voor alle goede feedback op mijn stukken, de begeleidingsgesprekken en je positieve instelling dat het proefschrift er zou komen. Het was erg fijn dat je deur altijd openstond en je altijd bereid was om mee te denken. Bedankt ook voor de fijne samenwerking, je was erg betrokken en deed altijd je best om snel feedback te geven. Zeker aan het einde van het traject heeft dit erg geholpen om de deadline te halen. Ik wil je ook bedanken voor de introductie in je (inter)nationale netwerk, bijvoorbeeld tijdens de door jou georganiseerde workshop en het bezoek aan Ofsted, wat mij de mogelijkheid gaf om een aantal vaak geciteerde auteurs in dit proefschrift persoonlijk te ontmoeten en mijn onderzoek met hen te bespreken. Je liet me zien dat ook de grote namen in ons vakgebied heel toegankelijk zijn en zo zaten we op een dag in het NH hotel mijn onderzoek te bespreken met John Hattie. Je stimuleerde me ook om mijn onderzoek aan een breed publiek te presenteren en zo presenteerde ik dit onder andere op de SREE in Washington aan mede-onderzoekers en op de ResearchED in Nieuwegein aan leraren. Allebei ontzettend leerzaam en waardevol voor mijn onderzoek.

Dankwoord

we de data niet op deze geavanceerde manier kunnen analyseren en je enthousiasme voor de gebruikte analyses was erg motiverend. Ik denk dat ons werk een mooi voorbeeld is van wat de samenwerking tussen psychometriërs en onderwijskundigen kan opleveren.

Hannah wil ik bedanken voor al haar inzet tijdens het project over de Impact! tool. Het was leuk en leerzaam om samen met Adrie en Bernard de eerste versie van de tool vorm te geven en samen het experiment op te zetten. Ik wil je ontzettend bedanken voor al je werk tijdens de dataverzameling, met name het werven van de leraren, het maken van de video-opnames en je werk als observator in mijn project. Ondanks dat de techniek ons regelmatig in de steek liet, lukte het toch om van alle leraren, verspreid over heel Nederland, drie bruikbare opnames te maken. Het verder ontwikkelen, promoten en onderzoeken van de tool ligt bij jou in erg goede handen en het is mooi om te zien hoe je verder vorm geeft aan dit project.

De reviewstudie was niet in deze vorm mogelijk geweest zonder alle informatie die instrumentontwikkelaars mij steeds hebben toegestuurd. Ik ben hen daar erg dankbaar voor. Tijdens de review hebben vijf beoordelaars met mij alle lesobservatie-instrumenten beoordeeld. Anja, Emmelien, Gerry, Linde en Martin, bedankt voor al het werk dat jullie verzet hebben voor de beoordelingen van de instrumenten. Anja en Gerry, ook bedankt voor jullie interesse en betrokkenheid tijdens mijn hele onderzoek. Jullie waren altijd bereid om mee te denken, bedankt daarvoor.

De dataverzameling voor het Impact! project was niet gelukt zonder al het gereis van Jet en Stef; bedankt voor jullie inzet bij het maken van de video-opnames. Ook alle 25 leraren die bereid waren om mee te werken aan het onderzoek, wil ik erg bedanken voor hun inzet en gastvrijheid, evenals hun 608 leerlingen die hebben meegewerkt. Linda wil ik bedanken voor het beoordelen van alle video’s. Je motivatie om dit gezamenlijk zo goed mogelijk te doen was erg fijn.

Op de voorkant van het proefschrift staat een foto van mijn oog. Observeren doe je immers met je ogen en je moet ook goed observeren om te kunnen zien wat het is. Wen, bedankt voor het maken van de foto, wat nog niet zo gemakkelijk bleek te zijn. Dion, bedankt voor het lezen van het gehele proefschrift en je goede taalkundige suggesties en correcties. Susan, bedankt voor het meedenken over de titel. Ik wil alle leden van de promotiecommissie bedanken voor het lezen van mijn proefschrift en voor het met mij in discussie gaan over de inhoud hiervan tijdens de verdediging.
Dankwoord

Tijdens de afgelopen jaren zorgden veel collega’s voor steun, leuke momenten en gezelligheid. Mijn promotieonderzoek begon bij de vakgroep OMD. Ik heb veel geleerd van de OMD-collega’s tijdens colloquia en ik wil hen bedanken dat ik altijd binnen mocht lopen voor vragen of een praatje. Ongeveer halverwege mijn promotieonderzoek verhuisde ik naar de afdeling ELAN, een afdeling waar ontzettend veel projecten gedaan worden die mijn interesse hebben en waar ik me erg thuis heb gevoeld. Ik wil ook alle ELAN-collega’s bedanken voor de fijne tijd. Emmelien, Marieke en Trynke wil ik daarbij in het bijzonder bedanken. Vanaf het begin tot het eind stonden jullie altijd klaar voor mijn vragen en heb ik veel aan jullie gehad. We hebben samen lief en leed gedeeld, en niet te vergeten ook veel verhalen over onze zwangerschappen en baby’s. Ik wil ook Nathalie heel erg bedanken voor alle gezellige treintijd, gezamenlijke werkdagen in Nijmegen en je bereidheid om altijd mee te denken, of het nu ging over de methode van mijn artikel of de kleur in mijn proefschrift.

Het werken aan mijn promotieonderzoek was vaak behoorlijk enzaam, iets wat niet heel goed bij me past. Het was daarom fijn als ik in het gezelschap van anderen kon werken. In het begin was ik daarom ook nog regelmatig bij de inspectie te vinden. Ik wil alle oud-collega’s bij Kennis bedanken voor hun aanmoediging, interesse en gezelligheid, zeker tijdens het eerste jaar van het promotietraject. Ik wil mijn kamergenoten Marjan bij OMD en Hannah bij ELAN bedanken voor het fijne gezelschap en ik wil ook iedereen op kamer 5436, (Anne, Christel, Elkse, Jitske, Rilana en in herinnering Wilma) bedanken dat ik in het laatste half jaar steeds bij jullie mocht aanschuiven. In het laatste jaar heb ik in Lent een werkplek gehuurd en ik wil iedereen van de Verdieping bedanken voor de fijne plek en gezelligheid.

During my PhD project I had the opportunity to also work on two publications not included in my dissertation. I would like to thank David from University College London and Herman from the Dutch Inspectorate of Education for the collaboration on a book chapter about feedback by Dutch inspectors to teachers. I would like to thank Courtney from ETS and Kirsti from the University of Oslo for the collaboration with Adrie and me on a publication about the qualities of classroom observation systems. Working on this publication really helped my thinking about classroom observation and therefore had a positive influence on this dissertation. Thank you also for being an inspiration for me to work hard on this interesting topic.
Dankwoord

In dit dankwoord wil ik ook stil staan bij de fijne mensen die ik privé om me heen had tijdens mijn promotie. Te beginnen bij de fijne groep **buren**, die vaak interesse toonden in mijn onderzoek en altijd klaar stonden voor een helpende hand. Ik wil **Mieke** heel erg bedanken, want door jouw trouwe hulp was het de afgelopen jaren altijd fijn om thuis te komen.

Mijn lieve vrienden wil ik bedanken voor alle interesse, support en bovenal gezelligheid. **Annelies, Inge, Machteld en Marleen** bedankt voor alle fijne etentjes op dinsdag, waarbij jullie mij ook de nodige support hebben geboden. Al bijna 10 jaar lang zien wij elkaar en ik hoop dat we deze traditie nog heel lang blijven voortzetten. **Alex & Lot**, wat fijn om beviend te zijn met jullie. Bedankt voor alle leuke vakanties en weekendjes weg in de afgelopen jaren, waardoor we even helemaal konden ontspannen. **Jody** bedankt voor alle kopjes thee en dat ik altijd bij jou terecht kon. Ik vind het fijn dat we al zo lang vriendinnen zijn. **Annelies en Lot**, ook bedankt dat jullie mijn paranimfen willen zijn, en **Jody** alvast bedankt voor het maken van de foto’s op 22 februari.


Dankwoord

Tot slot prijs ik me ontzettend gelukkig met twee fantastische mannen in mijn leven. Jasper, de afgelopen twee jaren waren onbeschrijfelijk geweldig. Ik ben ontzettend blij dat je er bent. Mochten er onverhoopt toch nog ergens gekke tekens in dit proefschrift staan, dan heb je ook nog een inhoudelijke bijdrage geleverd. Hier heb je samen met Cera (de poes) in ieder geval wel je best voor gedaan. Lieve Martijn, wat ben ik je ontzettend dankbaar voor alles wat jij in de afgelopen jaren hebt gedaan. Je was niet alleen op het werk adviseur, maar was ook die van mij. Ik heb veel van je geleerd over onder andere doelen en planningen, communiceren en het omgaan met lastige situaties. Ik denk niet dat het zonder jou was gelukt, je ondersteunde me op alle mogelijke manieren. Nu stimuleer je me om goed na te denken over wat ik wil in de toekomst en wat het beste bij me past. Bedankt dat je er voor me bent en voor het fijne thuis dat we samen hebben!
ICO dissertation series
ICO dissertation series

In the ICO Dissertation Series dissertations are published of graduate students from faculties and institutes on educational research within the ICO Partner Universities: Eindhoven University of Technology, Leiden University, Maastricht University, Open University of the Netherlands, University of Amsterdam, University of Twente, Utrecht University, VU University Amsterdam, and Wageningen University, and formerly University of Groningen (until 2006), Radboud University Nijmegen (until 2004), and Tilburg University (until 2002). The University of Groningen, University of Antwerp, University of Ghent, and the Erasmus University Rotterdam have been ‘ICO ‘Network partner’ in 2010 and 2011. From 2012 onwards, these ICO Network partners are full ICO partners, and from that period their dissertations will be added to this dissertation series.

343. Hubers, M.D. (08-09-2016). Capacity building by data team members to sustain schools’ data use. Enschede: University of Twente.


345. Scheer, E.A. (23-09-2016). Data-based decision making put to the test. Enschede: University of Twente.


350. Oonk, C. (07-12-2016). Learning and Teaching in the Regional Learning Environment: Enabling Students and Teachers to Cross Boundaries in Multi-


The quality and qualities of classroom observation systems

Marjoleine J. Dobbelaer