

University of Groningen

Unravelling the challenges of the data-based approach to teaching improvement

Helms-Lorenz, Michelle; Visscher, Adrie

Published in:
School Effectiveness and School Improvement

DOI:
[10.1080/09243453.2021.1946568](https://doi.org/10.1080/09243453.2021.1946568)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2022

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):
Helms-Lorenz, M., & Visscher, A. (2022). Unravelling the challenges of the data-based approach to teaching improvement. *School Effectiveness and School Improvement*, 33(1), 125-147.
<https://doi.org/10.1080/09243453.2021.1946568>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.



Unravelling the challenges of the data-based approach to teaching improvement

Michelle Helms-Lorenz & Adrie J. Visscher

To cite this article: Michelle Helms-Lorenz & Adrie J. Visscher (2022) Unravelling the challenges of the data-based approach to teaching improvement, School Effectiveness and School Improvement, 33:1, 125-147, DOI: [10.1080/09243453.2021.1946568](https://doi.org/10.1080/09243453.2021.1946568)

To link to this article: <https://doi.org/10.1080/09243453.2021.1946568>



© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 29 Jun 2021.



Submit your article to this journal [↗](#)



Article views: 1139



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 1 View citing articles [↗](#)

Unravelling the challenges of the data-based approach to teaching improvement

Michelle Helms-Lorenz ^a and Adrie J. Visscher ^b

^aDepartment of Teacher Education, University of Groningen, Groningen, the Netherlands; ^bELAN, Department of Teacher Development, University of Twente, Enschede, the Netherlands

ABSTRACT

The goal of this article is to clarify and unravel the complexity and challenges of improving teaching quality, based on measuring teaching quality and feeding back the results to teachers. We analyze different conceptualizations of teaching quality, and synthesize a framework for conceptualizing teaching quality in educational practice. We explain the pros and cons of four types of instruments for measuring teaching quality. Next, we scrutinize the requirements of effectively feeding back teaching quality data and the requirements for effective actions to improve teaching quality. We conclude with implications for improving the consequential validity of teaching quality measurements.

ARTICLE HISTORY

Received 20 December 2019

Accepted 15 June 2021


KEYWORDS

Teaching quality; data-based improvement; consequential validity

Introduction

Using teaching quality measurements to improve the quality of teaching has been common practice around the world for many years, and millions are yearly spent on related improvement-oriented activities. When one starts to think about these attempts, about what it encompasses and presupposes in terms of expertise, instruments, resources, and effort, it proves to be a really complex and demanding enterprise. The goal of this article is to analyze the challenges of measuring teaching quality for *formative* reasons, as well as the complexity of the processes involved in improving teaching quality based on teaching quality measurements.

When measuring aspects of teaching quality for formative reasons, the more or less explicit assumption is that the results of the teaching quality measurements are fed back to teachers who subsequently gain new insights about their teaching, and may use the feedback and the feedforward, if provided, to practice and subsequently improve their teaching quality. Teachers can either try to improve professionally on their own or participate in professional development trajectories. The intention is to improve teaching practice in the classroom and to ultimately improve student achievement (cf. Bell, 2012). However, the connection between measuring teaching quality and improving teaching quality is often not straightforward (e.g., Hu & van Veen,

CONTACT Michelle Helms-Lorenz  m.helms-lorenz@rug.nl

© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

2020a, 2020b). Various factors influence the measurement, how the measurement results become available to teachers, coaching approaches, and how teachers deal with the feedback and coaching for improvement purposes. All factors together will influence the extent to which teaching quality and student learning will improve (the term *consequential validity* is used in this context; Messick, 1989; Shepard, 1997).

Figure 1 presents an overview of the interrelated components of the formative process of measuring and improving teaching quality and the contextual factors influencing those processes. The model is not meant as a prescriptive linear model or as a description of how things work, but more as a logical model of what improving student outcomes based on measurements of teaching quality logically entails. Some processes may occur unconsciously and implicitly (e.g., conceptualizing teaching quality and working on improvement) compared to others (e.g., measuring teaching quality). In some cases measuring teaching quality leads to feeding back the results only, nothing more. Under the right circumstances it may lead to all subsequent steps, including the last two blocks in Figure 1, which depict the *intended effects* of formative data-based teacher improvement. The model applies to situations in which teaching quality data are provided to teachers by school-external entities as well as to situations in which teachers themselves (self-evaluation) or their peers, principals, or students provide them with information about their teaching quality. The whole process can vary from very top-down (teachers receiving feedback and executing improvement-oriented activities) to very

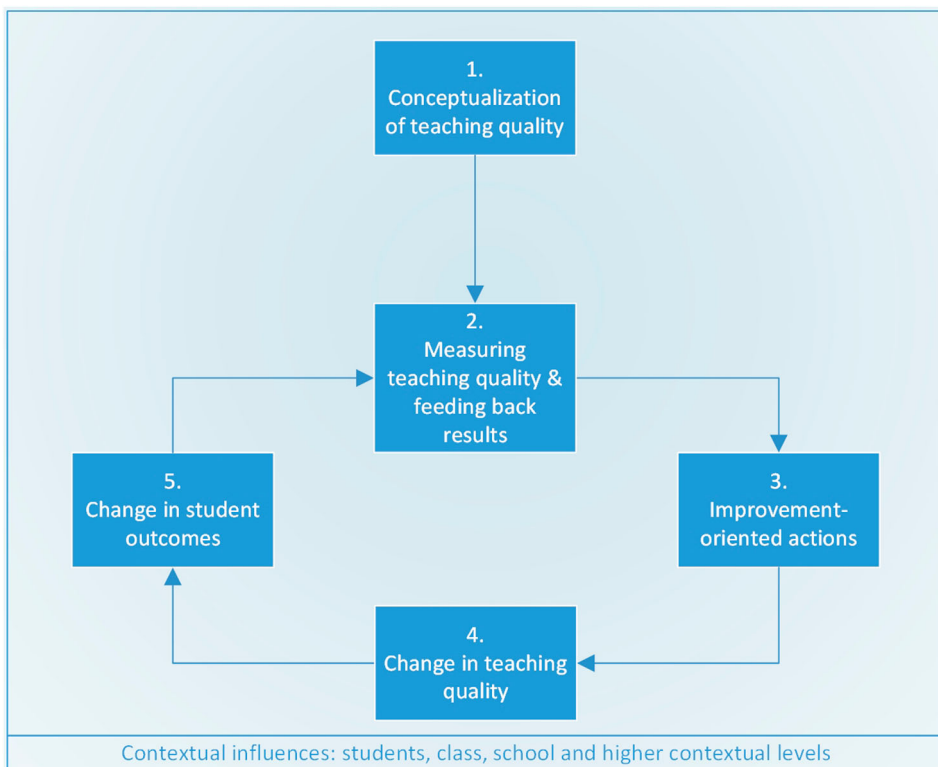


Figure 1. A step-wise model to unravel the complexity of data-based teacher improvement.

bottom-up (teachers anticipating how they conceptualize teaching quality and anticipating, or experimenting on, what the consequence of their behavior is on student learning). The outcome depends on the (un)conscious choices made in the intended improvement process. This article aims to analyze which factors influence improvement effectiveness when using teaching quality data and feedback/feedforward.

Figure 1 shows that if one aims to measure and improve the quality of teaching, the first question is how *teaching quality is conceptualized* (Block 1 in Figure 1). The depth and width of the conceptualization will impact the series of actions thereafter, and their results depicted in Figure 1. If teaching quality has been defined more or less explicitly and clearly, the second important question to be answered is *how it will be measured* (Block 2). This leads to a cascade of choices to be made in terms of, among others, the kind(s) of instrument(s) with specific psychometric qualities that will be used (e.g., lesson observations, and/or student questionnaires and/or student tests), who will rate teaching quality (e.g., principals, external consultants, students), and what the measurement conditions will be (Bell et al., 2019). For instance, will lessons be observed on the spot or be recorded and observed and rated later, will teachers themselves collect student opinions on teaching quality, or will this be done by a colleague (the latter may lead to more objective information), at what moment in time during the school year and the school week will the data be collected, and how many measures will be collected (e.g., Hill et al., 2012)? The choices made will affect the nature and the quality of the obtained teaching quality information, and the subsequent blocks depicted in Figure 1.

A second element of Block 2 involves *feeding back the results of teaching quality measurements* to teachers. Once teachers have been provided with information concerning their professional strengths and weaknesses, the nature and quality of the *improvement-oriented actions* (Block 3) taken by them and/or by relevant others (peers, principals, external coaches) will be decisive for the extent to which *teaching quality improves* (Block 4) and how that again influences *student outcomes* (Block 5). For the number and kind of improvement-oriented actions taken, the following questions are important. Is the feedback credible for the teachers? Which aspect(s) of teaching quality do teachers decide to work on (some aspects may be easier to improve than others)? Do teachers work in isolation to improve their teaching, or in cooperation? To what extent are they supported by their leaders, peers, and/or external coaches? Are the teacher professionalization activities grounded in psychological learning theory (to increase the possibility of improvement to occur)?

Figure 1 also includes a number of contextual factors influencing data-based attempts to improve teaching quality. For example, students' achievement levels may make it easier or more difficult to improve teaching quality. It may be easier to improve teaching quality in small, homogeneous classes than in large heterogeneous classes (Simpson, 2018). Features of the *school organization* teachers work in also matter (e.g., support for improvement, instructional leadership, the school performance culture), as well as the characteristics of the *higher levels* of the educational system teachers work in (e.g., the school district, school board, ministry of education, the school inspectorate; how much autonomy does each actor have for example with regard to teaching quality?). These contextual factors can make improving teaching quality more or less difficult.

The first three blocks in [Figure 1](#) will be elaborated now, by explaining the kinds of choices that can be made, and how those choices may impact the rest of the formative process.

Block 1: conceptualization of teaching quality

Teaching quality is considered to be the most important malleable educational condition rooted in the concept of *educational effectiveness* (= the “net” effect of malleable educational conditions on output; Scheerens, 2016). Scheerens’ (2016) over-arching multilevel model of educational effectiveness postulates *structured teaching* to entail more than actual teaching; a distinction is made between the pro-active, interactive, and retro-active aspects of teaching. The pro-active aspects concern all the preparation activities and prerequisites involved *before* a lesson is executed. Interactive aspects refer to the ([in]visible) interactions *during* the lesson. Retro-active aspects concern the evaluation of the conducted lesson and of student learning *after* the execution of the lesson, giving input for follow-up pro-active teaching.

Manifest teaching behavior in the classroom is viewed as a *proxy* of teaching quality reflecting the pro-active, interactive, and retro-active aspects of teaching. Studies on classroom teaching behavior in relation to (cognitive) student achievement have been reviewed, and second-order studies reveal more or less stable, visible effective teaching behaviors associated with greater levels of student learning gains (Creemers, 1994; Ellis & Worthington, 1994; Hattie, 2009; Levine & Lezotte, 1995; Marzano et al., 2008; Purkey & Smith, 1983; Sammons et al., 1995; Scheerens et al., 2005; Walberg & Haertel, 1992). These studies reveal at least six observable components of classroom teaching behavior showing a relationship with students’ learning and outcomes: Providing a Safe and Stimulating Learning Climate, Efficient Classroom Management, Clarity of Instruction, Activating Learning, Adaptive Teaching, and Teaching Student Learning Strategies. The theoretical foundation underlying the six components of classroom teaching was synthesized from teacher effectiveness research (e.g., Creemers, 1994; Sammons et al., 1995; Scheerens, 1992), combined with the scientific literature on learning environments and teacher support (e.g., Maulana, 2012; Maulana et al., 2013). Teacher–student interpersonal relationship has been shown to be an important determinant of the learning processes of students (den Brok, 2001; van Tartwijk et al., 1998). Comparable lists of effective teaching domains are found in Danielson (2013), Ferguson (2012), Muijs and Reynolds (2001), Pianta and Hamre (2009), and Scheerens and Bosker (1997).

Many teaching activities are observable *during the lesson*, but others are invisible, such as the teacher’s metacognitive activities during the execution of the lesson: keeping the (individual) learning goals in mind, orienting and sensing students’ states of mind (Strauss, 2005), monitoring goal-oriented activities, judging and interpreting, adjusting plans according to ongoing processes, and reflecting on the achieved goals to improve future lessons (reflection in action; Schön, 1988; Ward & McCotter, 2004).

The quality of teaching is influenced by various teacher characteristics as well as by situational factors. Teacher background characteristics that play a role are the professional training teachers have had and receive, and their professional experience as a teacher. These factors influence teachers’ professional competences: their *knowledge bases, professional skills, and attitudes*.

Baumert et al. (2013) refer to three teacher *knowledge bases*: (1) knowledge of the instructional potential of tasks, referring to local knowledge of tasks and multiple solutions, orchestration of tasks into instructional sequences, and the cognitive demand of tasks; (2) knowledge of creating meaning in interaction, referring to multiple representations and explanations, cognition of representations, fast recognition of mistakes, making use of critical incidents (maintaining the level of cognitive complexity, keeping students responsible for learning); (3) knowledge about students' conceptions and their thinking. Besides such subject-specific knowledge, teachers also have more general knowledge bases regarding student assessment, general didactical strategies, and knowledge of how students learn (best). Gess-Newsome (2015) defines six knowledge bases: assessment knowledge, pedagogical knowledge, content knowledge, knowledge of students, curricular knowledge, and topic-specific knowledge.

As teachers vary in quality (Haertel, 2013; Hanushek & Rivkin, 2012; Nye et al., 2004), as expressed in how much their students learn, teachers' *teaching skills* must also differ. Referring to the characteristics of effective teaching that were mentioned in the previous paragraphs, teachers differ for example in their classroom management skills, in how well they explain subject matter to students, and in the skill to adapt their teaching in line with student needs (e.g., Dobbelaer, 2019).

Teachers' *attitudes* have also been identified to influence (amplify and filter) teaching quality (Gess-Newsome, 2015), for example, teacher motivation and teacher morale (Troman & Woods, 2001), the passion for teaching (Day, 2004), professional beliefs as a result of age and career phase (Day, 2008; Fessler & Christensen, 1992), general professional self-efficacy (Rosenholtz, 1989), and more specific teaching efficacy beliefs (Tschannen-Moran et al., 1998).

Gitomer and Bell (2013) argue that students and teachers influence each other reciprocally while interacting with the content to be taught and learned. They postulate that teaching quality is determined by both parties together, by means of co-construction. Teaching quality not only depends on teacher knowledge, practices, and beliefs but also on the knowledge, practices, and beliefs of the students they teach, for example, the prior knowledge and skills and their motivation to participate in and contribute to lessons and to learn.

Teaching does not take place in one and the same standardized situation, not even in countries with a national curriculum. It is influenced by contextual factors such as class size and class composition (e.g., the degree of heterogeneity in terms of the socioeconomic status [SES] and performance levels of the students). The school organization also has its influence on the classroom; for example, the degree to which the school culture and school leadership demand and support quality teaching and the amount of available resources for teacher professional development. Finally, the domain of teaching is intertwined with critical larger context features such as a national curriculum and district and national level policies (e.g., Bell et al., 2019).

A summary is provided in Figure 2, and includes (non-exhaustive) examples and illustrates that teaching quality can be conceptualized in many different ways, for example, by focusing more or less on:

- what happens in the classroom (interactive teaching), or including the preparation (pro-active) and/or evaluation and reflection (retro-active) stages;

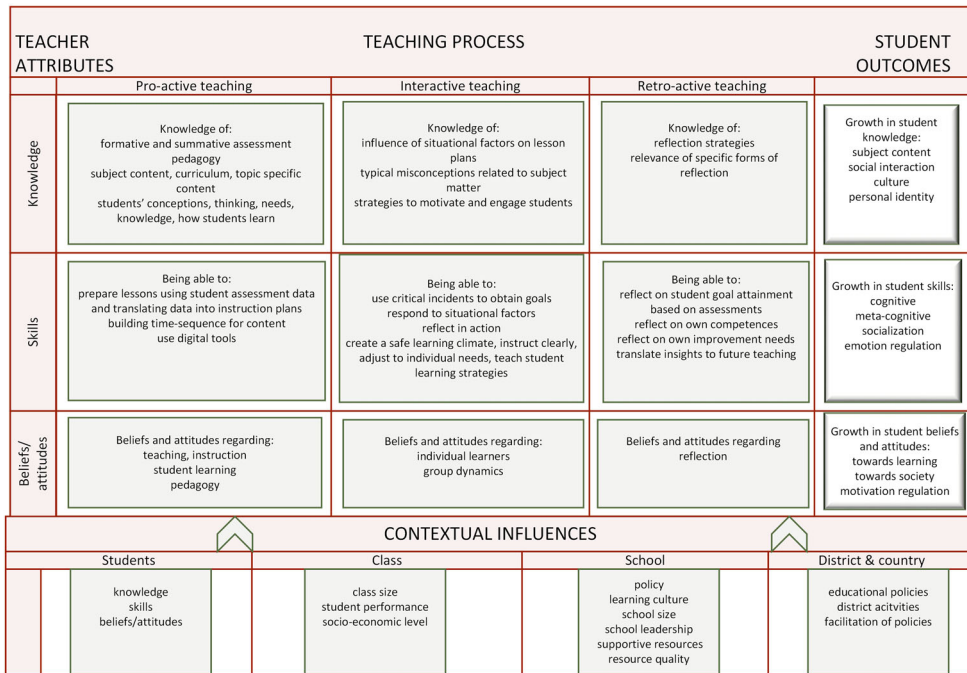


Figure 2. Conceptual model of teaching quality and factors influencing teaching quality.

- the teacher, or the student, or both;
- teachers’ skills and/or knowledge, and/or their beliefs/attitudes, or the knowledge, skills, and attitudes of their students, or one or more of these competences of both teachers and students.

The implicit or explicit conceptualization of teaching quality sets the stage for the consequential steps towards improvement, and is influenced by contextual student, class, school and supra-school level factors.

In Block 2 below, we will elaborate on how teaching quality, as conceptualized, more or less deliberately (e.g., by the choice for a specific lesson observation instrument) can be measured.

Block 2, Part 1: measuring teaching quality

The characteristics and the pros and cons of four different approaches to assessing teaching quality for formative purposes will be discussed now: (1) The measurement of teacher value added (VAM); (2) Teacher self-evaluation; (3) Lesson observation; (4) Student perceptions of teaching quality. In our view, each of these (and potential other) approaches should be judged using the following criteria (Trochim, 2006):

- *Practical feasibility*: the required resources (time, training, money) as well as the burden the measurement puts on teachers;
- *Face validity*: the extent to which the feedback is viewed by teachers, as covering the concept it claims to measure;

- *Reliability*: the internal consistency of scales measuring a teaching quality construct, and the stability of measurements over a specific period of time (e.g., within a week, measured under the same circumstances);
- *Construct validity*: the extent to which the instrument measures (some aspect of) teaching quality;
- *External validity*: the degree to which measurement results allow for generalizations to a larger universe of measurement results;
- *Predictive validity*: the extent to which the teaching quality measure is correlated with student achievement.

Re 1: teacher value-added measures

Teaching quality is conceptualized by this line of research as the degree to which teachers add value (VAM) to the starting achievement levels of their students (or, if a pre-test is not available, how much teachers perform above average in terms of the achievements of their students whilst correcting for relevant factors). McCaffrey et al. (2003) mention two reasons why VAM seems promising: (a) for separating the effects of teachers and schools from the powerful effects of such noneducational factors as family background, and (b) if these differences can be substantiated and causally linked to specific characteristics of teachers, the potential for improvement of education is assumed to be of great value.

In some countries it is relatively easy and cheap (and thus practically feasible) to compute teacher VAMs, if student test results are stored in databases that can easily be linked with teacher databases. If this is the case, it will often only be for some grades and for some subjects. If the student test and teacher data are not available, it will be much more difficult and expensive to apply this method. Because of the complex statistical models used for computing VAMs, it is difficult for practitioners to grasp what teacher VAM scores precisely indicate about the quality of their teaching (face validity).

When measuring value added, the big challenge is to isolate and validly measure the impact teachers have on the learning of their students amidst all other factors influencing student performance (Haertel, 2013). As classes and teacher-class combinations are not composed on the basis of randomization, the characteristics of classes and of teacher-class combinations (e.g., average classroom SES, teacher popularity) influence student learning besides the influence of a teacher's teaching quality. The number of factors to correct for is large; for example, the school features influencing teaching quality (e.g., teacher support, resources, and school climate, the quality of peer teachers), class size, class SES heterogeneity, and the level of class performance. It also will be very difficult to obtain valid information on all influencing factors, and the omission of one or more important factors cannot be ruled out. Moreover, the value added by teachers can be computed in many different ways (regarding how to weigh and correct for the included factors) leading to different models with different outcomes (McCaffrey et al., 2003). To date, there is no generally accepted, best teacher value-added measurement model, and thus VAMs' construct validity is not convincing (yet) (Timmermans, 2012).

Although VAMs are appealing for accountability purposes in some countries (the USA), their reliability and external validity can be questioned as experienced teachers' value-added scores prove to differ considerably between school years (Darling-Hammond,

2015). Especially for those teachers who are in the middle of the performance distribution (neither very good nor very poor teachers), reliability is low. One way to improve the stability and external validity of teacher VAM could be to enlarge the number of measurements by using the student achievement data of teachers' students over several school years.

In the Measures of Effective Teaching (MET) project (Bill & Melinda Gates Foundation, 2018), not surprisingly, prior achievement of teachers' students proved to be the best predictor of those teachers' future students' outcomes (predictive validity). It is quite likely that the same applies to teachers' VAMs.

Re 2: teacher self-evaluation

It is not rare that teachers evaluate their own teaching quality. In line with other professionals (e.g., doctors, lawyers), the core of their professional work is fairly unpredictable and therefore can only be standardized to a limited degree (Mintzberg, 1979). Teachers have considerable professional autonomy. They have been trained for many years and are supposed to be able to deal with complex and unpredictable professional situations, and also to be able to evaluate and improve their professional functioning. For medical professionals, Eva and Regehr (2005) illustrated that this assumption is "overly optimistic" because of the "*Lake Wobegon effect*" (Kruger & Dunning, 1999). Kruger and Dunning (1999) showed that most of us think that they perform above average. Especially lower performing individuals in a specific domain lack the metacognitive skillfulness for valid self-assessments. The researchers argue that "... the skills that engender competence in a particular domain are often the very same skills to evaluate competence in that domain – one's own or anyone else's" (p. 1130). Thus, the stronger the need to improve professional skills, the weaker the awareness is of the need to improve. Top performers prove to underestimate their performance, probably because they know best what excellent performance looks like and how much it requires. The Dutch Inspectorate (Inspectie van het Onderwijs, 2013) found that almost all Dutch teachers had a positive image of their own teaching skills, especially of their basic teaching skills. According to the Inspectorate, 35% of secondary school teachers and 66% of primary school teachers had an accurate perception of their professional skills. Dobbelaer (2019) studied how teachers, their students, and external observers rated the same lessons given by that teacher using the same items, and she found that teachers were most positive, students somewhat less, and external observers the least positive about the teaching quality in those lessons. Gitomer et al. (2014) also found that teachers rated themselves higher on the Classroom Assessment Scoring System (CLASS) observation protocol dimensions, compared to observers. The validity of the ratings may differ between aspects of teaching quality. Agreement between teacher self-ratings and observer ratings was high for classroom organization, modest for emotional support, and non-existent for the instructional support of teachers. Teachers rated themselves much more positive on instructional support and emotional support and less positive on classroom organization compared to observers.

Teacher self-evaluation surveys also vary widely in how they conceptualize and operationalize teaching and thus also vary in construct validity, which will impact the external validity and predictive validity of the ratings.

Thus, although teacher self-evaluations of their teaching may be relatively easy to conduct (practical feasibility) and also have face validity for practitioners, they may not provide a valid starting point for improving teaching. Especially in the case of low- and top-performing teachers (even if the ratings are reliable over time), self-evaluations lack construct validity, external validity, and predictive validity.

Re 3: lesson observation

An important strength of lesson observation results is its face validity, as the feedback based on a lesson observation clearly refers to the core of the teaching job (quite different from teacher VAM scores). Observing lessons is quite popular in educational practice but more difficult to conduct well than often realized (Bell et al., 2019). First of all, it requires valid lesson observation instruments (LOBs) that meet a number of criteria. The teaching constructs in the instrument ideally reflect teaching behaviors that have a scientifically proven, positive relationship with student learning, or that matter for some other reason (e.g., students simply should feel safe in class, and teaching should match students' needs). The instruments also should include items that validly operationalize the constructs measured. This may not be easy as our knowledge of what quality classroom management, instructional differentiation, cognitive activation of students, and so forth, look like is limited. Dobbelaer (2019) conducted a worldwide review of LOBs for primary education and found that the empirically proven reliability, construct validity, and external validity of the LOBs vary considerably and that for many instruments there is substantial room for improvement.

Scholars also argue that for statistically reliable measures of teaching quality a minimum number of lessons of a teacher should be rated. Hill et al. (2012) revealed that reliable teaching quality scores as measured by means of the Mathematical Quality of Instruction (MQI) lesson observation instrument requires three to four lessons to be observed and rated by three to four raters. This finding was confirmed in other contexts (e.g., van der Lans et al., 2016). In practice, multiple observations are time consuming and expensive and thus not that practically feasible. The required number of lesson observations depends on the intended claims to be made on the basis of the observations. If the goal is to estimate the average teaching quality of a single teacher and its consistency (e.g., the average quality of the approximately 1,000 lessons a single teacher for primary and secondary education in the Netherlands delivers during a school year), then many different lessons at different time points in the school year (different days, lesson periods, school year phases) will be needed. If the observation goal is to obtain an impression of a single teacher's teaching quality at a certain point in time, then rating multiple lessons in a short period of time by multiple raters might do. Observation instruments used for individual assessments that lead to high-stake decisions or consequences should meet the highest psychometric requirements. If instruments are used to generate formative feedback aiming to improve the teaching quality, a single observation can suffice if the teacher (and/or the teacher's students) can confirm the representativeness of the observed lesson for that period of the year. If not, then more observations will be needed. It also makes a difference who observes and rates. For example, it may be harder for a principal or a peer teacher to rate a peer teacher objectively, compared to an external professional observer. In the USA it was found that 99% of

the teachers observed by their principals were rated “good” or “great” (the Widget effect; Weisberg et al., 2009). Reliable and valid lesson observation scores require raters to be well trained and certified (and their observation skills should be calibrated periodically on a regular basis).

As far as the predictive validity of lesson observation scores is concerned, relationships between teachers’ scores on observation instruments and student achievement outcomes generally range from weak to moderate (moderate is 0.3–0.4). The MET project (Bill & Melinda Gates Foundation, 2018) showed that classroom observations based on a number of widely used American LOBs added somewhat to the predictive validity of (earlier) student achievement measurements (the best predictor of follow-up student achievement).

Effective teaching is considered to be situational to some extent (e.g., Bell et al., 2019). The question is therefore whether we can capture situational effectiveness with standardized observation instruments. Is the influence of the students also measured and accounted for? It is easier to teach well in some classrooms than in others because teaching and learning are also a matter of co-construction between students and teachers. If students are more eager to learn and to participate in a constructive way, then it will be more probable for a teacher to score well on lesson observation standards. One could argue that lesson observation scores should be corrected for situational features influencing teacher practice, for example, class size, class SES, within-class student achievement variation, the school performance culture, and available teaching resources. This will not be easy to do at all. Instead, it is probably more feasible to use the knowledge and information about the specific situation measured in the feedback that is given to teachers, to allow for more situational improvement plans (Hu & van Veen, 2020a).

In sum, the reliability, external validity, construct validity, and predictive validity of lesson observation instruments varies; thus, it is important to pay careful attention to their qualities when choosing a LOB. This may, however, not be easy for practitioners: what is important to look at, where to find the relevant information, and how to weigh all available information? Lesson observations conducted in accordance with all the prerequisites that have been discussed here are labor intensive and expensive, which negatively affects their practical feasibility. As a result of the unawareness of what conducting quality lesson observations takes, the prerequisites for valid lesson observation are quite often violated in educational practice. School-wide lesson observations conducted by external experts in many cases will be too expensive for many schools. Lesson observations are strong in terms of their face validity. If the feedback in terms of the lesson observation scores is valid, then it can be a rich start of an improvement process; however, as said, it will not be easy to accomplish the prerequisites to be fulfilled in educational practice.

Re 4: student perceptions of teaching quality

Measuring teaching quality by means of student perceptions of teaching quality (SPTQs) is gaining popularity (e.g., Bijlsma et al., 2019; Ferguson, 2012; van der Lans et al., 2015; van der Scheer et al., 2019). Students represent the target group, the schools’ “clients”. In many other organizations, the client perspective is crucial for maintaining and improving client service. Compared to lesson observations, student perceptions can be

measured easily and at low cost, especially now that digital instruments have become available for measuring student perceptions, for automatically storing and reporting the collected information, and for distributing the results to teachers (Bijlsma et al., 2019).

Just like any other instrument for measuring teaching quality, SPTQs are imperfect. With respect to their construct validity, a number of things are important. Students potentially may be the best, most valid source of information for particular aspects of teaching (e.g., how they personally experience the teacher's interaction with them, their instruction, and expectations). They may be less capable of rating other aspects of teaching, for example, how correctly teachers use mathematical concepts, or how well they promote student self-regulation. An important question is if students can differentiate in their ratings between various aspects of teaching quality. van der Scheer et al. (2019) found that Dutch Grade 4 students can.

SPTQs represent students' subjective opinions of the quality of teaching. Such subjective opinions may be biased because of teacher characteristics (e.g., the teacher's popularity, appearance, humor, and gender). Characteristics of the students themselves may also bias student ratings, for example, the extent to which a student likes the subject taught, the motivation to rate seriously, and student gender in relation to teacher gender. We do not know much yet about the extent to which the validity of student teaching quality ratings is influenced by such factors in concert (Bijlsma et al., 2020).

The same average SPTQs scores can reflect quite different interpretations. Students with different performance levels and instructional needs might value the same teacher's behavior differently. High-performing and low-performing students may, for instance, have quite different reasons for rating the pace of instruction negatively (e.g., too slow, too fast). As such, it is interesting to distinguish between how low-performing, average, and high-performing students assess teaching quality and why each of these groups does so.

SPTQs have the advantage that classes have a large number of raters (compared to 1 or 2 observers rating a lesson) rating simultaneously. This positively influences measurement stability/reliability (regression to the mean). Moreover, students can give their view on how the teacher teaches *in general* during a school year (this would require many lesson observations), which also enhances the reliability and external validity of SPTQs.

Not that much is known about the predictive validity of SPTQs, although den Brok et al. (2004) found that the nature of interpersonal student–teacher behavior as perceived by students explained up to more than half of the variance in student outcomes at the teacher-class level, and the MET project (Bill & Melinda Gates Foundation, 2018) showed that students' perceptions of teaching quality combined with students' pre-test achievement data and teaching quality ratings from trained observers were more predictive of follow-up student outcomes than pre-test achievement data and lesson observations scores only.

In summary, the collection of SPTQs is practically feasible and efficient as it nowadays can be done by means of digital devices. SPTQs may be a valuable source of information for teachers regarding the quality of their teaching especially if teachers discuss the feedback with their students. Because of the number of raters involved, SPTQs can provide statistically reliable information (not in very small classes). The extent to which SPTQs can be used for validly measuring (specific aspects of) teaching quality (construct validity) will depend on the quality of SPTQ questionnaires and on the extent to which students'

subjective opinions are biased. The constructs measured ideally have been proven to matter for student learning and can be rated validly by students. Students probably rate some aspects of teaching quality more validly than other aspects. The latter is important for the extent to which teachers believe that students' opinions about their lessons are valid (face validity). Some studies support the predictive validity of SPTQs, but this topic requires more research.

Wrapping up, we argue that VAMs seem to have little value for informing the improvement of teaching quality and teacher self-evaluations lack validity. The other two measures of teaching quality can be improved in various respects. In our opinion we should invest much more in improving the quality and feasibility of those instruments for evaluating the quality of the core activity of schools for formative purposes. We cannot rank the best to the worst performing teacher with full certainty and validity, but that also should not be our goal. Providing teachers with formative feedback based on measurements of the teaching process has proven to be a beneficial lever to the improvement of student achievement (Faber et al., 2017; van Geel et al., 2016).

Relating to the conceptual model of teaching quality in Figure 2 leads to the following observations with respect to how teaching quality usually is measured in schools: (1) Measurements mainly focus on interactive teaching and to a much lesser extent on value-added measures of student outcomes. (2) Pro-active and retro-active teaching receive scarce attention. (3) Student outcomes are primarily measured in terms of students' subject knowledge of the core school subjects. Other aspects of student knowledge, student skills, student beliefs, and attitudes are not measured frequently as indications of the outcomes of teaching. (4) Teaching skills are mainly measured in the classroom setting. (5) Teacher knowledge, teacher beliefs, and attitudes are rarely measured. (6) Correcting the measures for contextual influences is exclusively applied in teacher VAMs. (7) Teaching quality measures usually do not include subject-specific teaching behaviors. Table 1 provides an overview of our conclusions regarding the practical feasibility, reliability, and validity of four ways to measure teaching quality.

Block 2, Part 2: results feedback

Feedback has amongst the most powerful influence on learning and performance improvement (Hattie & Timperley, 2007; Kluger & DeNisi, 1996). Hattie and Timperley (2007) identify three major feedback-related questions: Where are you going? (feed-up), How are you going? (feedback), and Where to next? (feedforward). The four ways

Table 1. Practical feasibility, reliability, and validity of four measures of teaching quality.

	practical feasibility	face validity	reliability	construct validity	external validity	predictive validity
value-added measures	--	--	-	--	-	++
teacher self-evaluations	++	++	?	--	--	--
lesson observations	--*	++	+/-**	+/-**	+/-**	+/-**
student perceptions	++	+/-***	++	+/-	++	+/-

*If all requirements are to be met. **Will depend on how much the requirements are met. ***Can vary between teachers and cultures (in some cultures student voice is more welcome than in others).

to identify gaps between actual teaching quality (“How are you going?”) relative to, explicitly or implicitly, desired teaching quality (“Where are you going?”) that were discussed in this article can form the starting point for improvement action (“Where to next?”). Feedback can increase motivation or engagement and effort to reduce the discrepancy between where one is and where one would like to be. It can also lead to more cue searching and a better understanding of how things work and how things can be improved, which can lead to improved task processes. However, the impact of feedback is more complicated than meets the eye.

In their famous meta-analysis of feedback research, Kluger and DeNisi (1996) revealed that feedback interventions impact learning positively as well as negatively. Their feedback intervention theory (FIT) states that learning in response to feedback will depend on the cues provided in the feedback in combination with specific task characteristics, personality factors, and situational variables.

First of all, tasks vary in complexity and in the degree to which it is known what the most effective ways to execute a task are. Teaching is a very complex context-specific task, which makes it hard to diagnose and improve (under)performance in response to feedback, also because we do not have unambiguous instruments to measure teaching quality.

As far as the influence of the *feedback cues* are concerned, Kluger and DeNisi (1996) postulate that a negative discrepancy in the feedback information compared to a standard will most probably lead to increased effort. The effort will be maintained if it leads to a reduction of the discrepancy. However, if it does not lead to improvement, then the attention will shift from the task to the self, which will impede learning. Positive feedback most probably will lead to raising one’s own standards and increase effort, but it can also lead to keeping the same standards and even to reducing efforts (William, 2011).

In addition to the positive or negative sign of the feedback, other feedback characteristics also influence its effectiveness. Crucial is how much the feedback recipients can learn from the feedback about their performance, as well as to what extent the feedback supports performance improvement by giving hints about how improvement can be accomplished. Feedback characteristics that matter are (Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Visscher, 2015; Visscher & Coe, 2003):

- the comprehensibility of the feedback;
- the practically feasible feedback frequency: frequently enough to enable performance monitoring, but not too frequent, to prevent being a burden;
- the time laps between task execution and feedback provision (in case of much delay, the answer to “How am I going?” may already have changed too much);
- the content of the feedback in terms of whether the feedback is about *task* performance (performance is good or bad), the *process* (of task execution), *self-regulation* (directed at the monitoring and regulation of actions toward the learning goal), or the *self* (e.g., “Well done” or “That’s an intelligent response”). Feedback on the self can imply praise or be threatening and does not enhance learning (Hattie & Timperley, 2007);
- the degree to which the feedback only indicates whether overall performance is “good” or “bad” or (also) provides more fine-grained information regarding (a) particular task component(s) that can be improved;
- the degree to which the feedback indicates how performance can be improved.

Table 2. Feedback characteristics per measurement type.

	feedback comprehensiveness	practically feasible feedback frequency	time laps execution – feedback	feedback on task/process/ self- regulation/self	fine- grained	improvement hints
value-added measures	--	--	--	T/S	--	--
self- evaluations	++	--	++	P/S	+/-	--
lesson observations	++	-	++	P/S	++	++
student perceptions	++	++	++	P/S	+	+

Note: T = task; S = self; P = process.

Table 2 shows that we think that in most cases the results of the four teaching quality measurements as such (thus, apart from the interpretation of the scores), apart from value-added measures, will be quite comprehensible.

Student perceptions of teaching quality these days can be fed back frequently in a digital form (Bijlsma et al., 2019). The other forms of feedback in the practice of schools generally are infrequent because of the required resources and lacking information (VAM). Value-added measures of teaching quality in many cases become only available once or twice a school year, if at all, which is not beneficial for improving teaching quality (e.g., for monitoring the effects of the improvement measures taken). The other three measurement types can provide feedback without much delay. Whether the feedback points more to the task, process, or the self depends on what is measured, how the feedback is presented to teachers, and how the feedback is interpreted and attributed by a teacher (Hattie & Timperley, 2007): Are measurement results attributed to their efforts, or, for example, to the quality and efforts of their students, or incorrect judgments by lesson observers or students? This does not apply to teacher self-evaluations, but, as indicated, few teachers rate themselves as underperformers.

Lesson observations and SPTQs, especially if combined with a dialogue about the feedback between, respectively, teacher and lesson observer and between a teacher and their class, provide the most fine-grained feedback.

VAMs do not provide hints for individual improvement. Self-evaluations also do not inform teachers much regarding to what they should work on and how to improve professionally. Compared to lesson observations combined with coaching, SPTQs provide teachers with less information about how teaching quality can be improved. Students can be asked to also provide recommendations for improvement as part of the assessment, or after the results have become available, in a dialogue about the feedback between teachers and students. However, such improvement hints may be limited in terms of expertise, compared to improvement support from experienced coaches.

Feedback effects are not only dependent on the cues in the feedback but, in case the feedback is given by a person, also on characteristics of the *feedback giver*: for example, what do they consider important, how much teaching experience do they have themselves, do they show empathy and support towards the teacher (Wiliam, 2011).

The characteristics of *feedback recipient(s)* (Kluger & DeNisi, 1996, use the term “personality factors”) are also important for the effect the feedback will have. Does the recipient accept or reject the feedback, for example, the opinions of students about their teaching

quality (more about this in Block 3 below)? How motivated is a teacher really to improve? The latter may correlate with a teacher's age, experience, and self-efficacy and a teacher's performance level. The extent to which a teacher has the competences to improve in response to feedback will also influence the feedback impact (Visscher, 2021). Feedback to persons highly familiar with a task (e.g., experienced teachers) can be inhibiting because it interrupts automatic scripts. Feedback to persons less familiar with a task can lead to generating and testing personal hypotheses. If the hypotheses match reality and are deemed to be correct, they can lead to learning effects. If not, learning may not occur (Kluger & DeNisi, 1996).

Hu and van Veen (2020a) describe two coaching strategies based on feedback provided after lesson observations: prescriptive and collaborative coaching styles, in which the latter coaching pedagogy creates more constructive dissonance.

With respect to the *situational factors* influencing feedback impact, one can think of the following examples: the degree to which the school environment is safe (e.g., for openly discussing an attempt to improve performance), improvement-oriented, cooperative, and supportive and facilitates improvement in terms of the resources required for working on improvement (e.g., time and money; Schildkamp & Kuiper, 2010).

The empirical study of Hu and van Veen (2020b) revealed that meaningful teacher engagement in a professional development (PD) program, involving lesson observations and feedback, typically occurred in three interrelated conditions: (1) voluntary participation in the PD program; (2) a safe and collaborative PD culture, which allowed the dissonance to be constructive rather than destructive; and (3) the creation of sufficient dissonance between what teachers already know and the new information provided in the PD program (cognitive/conceptual friction).

Overall, we conclude that lesson observations and SPTQ generate feedback content that provides a good basis for formative purposes, and that the characteristics of the feedback giver, the recipient, and the context of the feedback recipient influence feedback effects.

Block 3: improvement-oriented actions

Feedback does not always reach the target group (Weiss, 1998), but if teachers, teacher departments, and/or principals do receive some form of feedback regarding teachers' professional strengths and weaknesses, then the nature and quality of the follow-up activities conducted by the recipient(s) and relevant others (e.g., peers, principals, external coaches) are decisive for the extent to which teaching quality (Block 4) and student outcomes (Block 5) will improve.

If feedback is not ignored or rejected, and teachers are willing to use some form of feedback for improving their professional competences, that still may be difficult (Wiliam, 2011). If a teacher attempts to improve professional competence, hopefully the focus will be on the right aspect of that competence. If not, then much effort may be wasted. As explained in Block 2, feedback may not always be fine-grained enough to know what to work on precisely. If the feedback is more detailed, and for example indicates insufficient teacher classroom management skills, or limited cognitive activation of students, then the improvement activities can be targeted to those skills. Such feedback still requires a further diagnosis of what precisely is below standard (which aspect of

classroom management, etc.) and what causes underperformance. Improvement then still may be difficult, for instance, because teachers may not have an idea of how aspects of their professional competences and performance can be improved, or, in case they do know this, they themselves may not have the skills or resources to accomplish improvement (Weiss, 1998).

How do professionals in other fields besides education improve? Ericsson's (2006) research of expert performance in domains like medicine, music, chess, and sports shows that expert performers in these domains acquire their superior performance by means of *deliberate practice*. They have a strong motivation to improve, deliberately step out of their comfort zone, search for those performance aspects that are not perfect yet, and focus their improvement activities on a small performance aspect where there is room for improvement. After formulating a very precise improvement goal, they work intensively on accomplishing the goal. This improvement work is done regularly and intensively, but each time only for a short period of time (e.g., every day for 30 min) as it is tiresome because they are not good at what they try to improve. The work is continued until the improvement goal has been accomplished.

Translated to our topic, this raises the question what quality teaching or specific aspects of quality teaching (e.g., quality instructional differentiation) look like in detail and which knowledge steers teachers' decisions. Such quality benchmarks are important for teachers who want to evaluate their performance. If such information is not available, solid knowledge about how teachers who do not (yet) possess specific desired teaching competences can acquire those competences effectively will be lacking too. If you cannot measure it, you cannot improve it (Ericsson, 2006). In our opinion, our knowledge base on quality teaching and how to get more of it is still limited. Some evidence on teaching characteristics that correlate with student achievement (e.g., Brophy & Good, 1986; Rosenshine, 2012) derived from process-product research is available. However, detailed, generally accepted knowledge of what all components of pro-active, interactive, and retro-active teaching (cf. Figure 2) ideally look like is not available. The same applies to how specific teaching skills (e.g., how to teach student self-regulation well) can be acquired best. Moreover, teacher professional development activities are not seldom undertaken without much attention for the learning-psychological prerequisites for professional learning (Kennedy, 2016).

The literature about designing training programs for acquiring all sorts of, not necessarily education-specific, complex professional competences is much more detailed and empirically validated than the literature on teacher professional development, and thus is something from which teacher professional development can benefit much. The four-component instructional design (4C/ID) model by van Merriënboer and Kirschner (2007) is a validated methodology for the cognitive task analysis of "quality" professional task execution (e.g., the analysis of how expert air traffic controllers, expert medical specialists, expert software designers, and expert teachers work and reason), and for subsequently designing professional training programs for acquiring the competences for quality task execution. Training programs are designed on the basis of the results of a cognitive task analysis, following an explicit learning-psychological rationale. In the training programs, learning tasks are included that are representative of the complex task to be learned. Learners start with a simple task and continue with more complex tasks if a task is mastered at a specific level of complexity. The learner

support given to a learner learning the tasks at first is elaborate but decreases as tasks are mastered better by learners. The learner is provided with information that is important for executing the routine and non-routine subtasks: the cognitive strategies and mental models that are important for taking decisions in varying situations (van Merriënboer & Kirschner, 2007).

Common practice is that teachers receive some form of feedback on their professional performance. Improvement-oriented actions in response to the feedback may, among others, include searching for relevant information in the literature and on the internet, attending a conference, observing lessons of colleagues (of the same, higher, or a lower professional quality), taking a course, receiving coaching from an expert or colleague, engaging in lesson study, and starting a professional community with peers. The effects of these improvement-oriented activities will be influenced by the extent to which:

- the improvement activities address a teacher's developmental needs: Did the recipient interpret the feedback correctly and choose a performance aspect that indeed requires improvement?;
- the characteristics of quality and expert performance are known for the teaching aspect one tries to improve;
- the designed professional development activities are based on a valid learning theory.

Bringing it all together

The aforementioned analysis has shown that the consequential validity of measuring teaching quality for improvement is far from self-evident. Working on the improvement of teaching quality in a data-based way is a very complex process that only will be effective if many preconditions have been fulfilled.

Teaching quality has many different aspects, and every teaching quality measure is imperfect in terms of reliability and validity. No single measurement approach captures the full range of the pro-active, interactive, and retro-active aspects of the teaching process, nor the range of teacher knowledge, skills, and attitudes required for the various stages of teaching. This implies that conclusions about the quality of a teacher's teaching always should be drawn with care. Neither does this imply that attempts to estimate proxies of teaching quality are useless, nor that measurement improvement is not needed. Being aware of the imperfections of our measures, we should deliberately work on improving them step by step, for example, by improving the operationalization of the constructs measured, by increasing the focus on low-inference items, by combining the perspectives of lesson observers with those of students, by incorporating and adjusting for relevant context information, and by incorporating (subject-specific) knowledge, skills, and beliefs of teachers in the operationalization of teaching quality. Our analysis revealed that lesson observations and SPTQ meet our evaluation criteria more favorably compared to VAM and teacher self-evaluations. We as researchers should investigate how the instruments for measuring the student and the observer perspective can be used in such a way that the strengths of each perspective can be benefitted from. Maybe measures of

teachers' value added could also be improved. The value teachers add on average to students' knowledge, attitudes, and skills, across many school years, might become an accurate indicator of the results of their teaching. Better ways to calculate teacher VAMs and to translate the VAM indicators into easily comprehensible information for teachers will also be needed, and VAM measures should cover more than students' academic performance in a few core subjects in the final grades (see [Figure 2](#) for examples of other relevant student outcomes).

Working on improvement in response to feedback presupposes a whole range of competences and resources: knowledge about how to interpret and deal with the feedback for improvement, a qualified feedback giver, a feedback recipient, and a school team that really strive towards improvement and that support and facilitate the improvement-oriented activities. In our view, it requires too much expertise and too many resources (e.g., time and money!) of the average school to be able to do this on their own given that they have not been trained for such complex activities. This means that for teachers to become "as good as possible" in their core work, and to improve during their whole career, their professional development will have to be organized in a different way. Schools need the time and financial resources to work in cooperation with external experts towards improving teaching quality as a regular part of their work. If standards are to be raised, other approaches to professionalization are required. If continuous development is the aim, then governments should spend much more money on education, such that teachers will have the time and facilities to work on improving complex teacher competencies in a profound way, by paying attention to pro-active, interactive, and retro-active teaching.

In addition, empirically verified standards for quality teaching (e.g., quality differentiation, quality classroom management, quality formative evaluation) need to be developed on the basis of empirical research. Such professional standards can be derived from in-depth cognitive analyses of how expert teachers in specific teaching aspects teach, think, and decide. Clearer conceptions of quality teaching will allow for better evaluations of teacher skills, knowledge, and beliefs. The complexity of acquiring complex teaching skills is often underestimated. Too many teacher professional development activities lack evidence-based, learning-theoretical foundations and as a result do not support the accomplishment of the learning goals adequately. The use of validated instructional design models like the 4C/ID model (van Merriënboer & Kirschner, 2007) can be of great value for designing better teacher professionalization interventions. Such designs should be tested and optimized, and along that road eventually lead to larger scale interventions for acquiring the standards for high-quality teaching (cf., Borko, 2004).

Changing and improving teachers' knowledge, attitudes, and skills in a data-based way is far from easy, changing teachers' practices in their classrooms is even more difficult, and improving the achievements of their students on a large scale is even harder. But we do have examples of successful attempts (Allen et al., 2011; Borman et al., 2007; van Geel et al., 2016). Building on those and taking account of the recommendations made here can bring us further regarding our knowledge base of how teachers can learn and continue to develop their competences in such a way that their students also learn and develop more.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Michelle Helms-Lorenz is an associate professor at the Department of Teacher Education, University of Groningen, the Netherlands. Her research interests include teaching skillfulness and wellbeing of beginning and pre-service teachers and effective interventions to promote professional growth and retention.

Adrie Visscher is a full professor at the University of Twente, the Netherlands. He is interested in measuring teaching quality validly as a basis for improving teaching quality. His research focuses on how teachers can be supported in optimizing the quality of their lessons and their impact on student learning by providing them with *feedback*: feedback about the features of their teaching activities (e.g., based on student perceptions or lesson observations) and feedback about their impact on student achievement. As this kind of feedback can be the starting point of improvement-oriented activities, his research also focuses on how teachers can be trained effectively for differentiating their teaching activities in line with students' varying instructional needs.

ORCID

Michelle Helms-Lorenz  <http://orcid.org/0000-0001-9314-6962>

Adrie J. Visscher  <http://orcid.org/0000-0001-8443-9878>

References

- Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science*, 333(6045), 1034–1037. <https://doi.org/10.1126/science.1207998>
- Baumert, J., Kunter, M., Blum, W., Klusmann, U., Krauss, S., & Neubrand, M. (2013). Professional competence of teachers, cognitively activating instruction, and the development of students' mathematical literacy (COACTIV): A research program. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Mathematics teacher education: Vol. 8. Cognitive activation in the mathematics classroom and professional competence of teachers* (pp. 1–21). Springer. https://doi.org/10.1007/978-1-4614-5149-5_1
- Bell, C. A. (2012, September 13). *Validation of professional practice components of teacher evaluation systems*. [Paper presentation]. 14th annual Reidy Interactive Lecture Series, Boston, MA, United States.
- Bell, C. A., Dobbelaer, M. J., Klette, K., & Visscher, A. (2019). Qualities of classroom observation systems. *School Effectiveness and School Improvement*, 30(1), 3–29. <https://doi.org/10.1080/09243453.2018.1539014>
- Bijlsma, H. J. E., Glas, C. A. W., & Visscher, A. J. (2020). *Factors related to differences in digitally measured student perceptions of teaching quality* [Manuscript submitted for publication]. Faculty of Behavioural, Management and Social Sciences, University of Twente.
- Bijlsma, H. J. E., Visscher, A. J., Dobbelaer, M. J., & Veldkamp, B. P. (2019). Does smartphone-assisted student feedback affect teachers' teaching quality? *Technology, Pedagogy and Education*, 28(2), 217–236. <https://doi.org/10.1080/1475939X.2019.1572534>
- Bill & Melinda Gates Foundation. (2018). *Measures of effective teaching project FAQs*. <https://k12education.gatesfoundation.org/blog/measures-of-effective-teaching-project-faqs/>
- Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher*, 33(8), 3–15. <https://doi.org/10.3102/0013189X033008003>

- Borman, G. D., Slavin, R. E., Cheung, A. C. K., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2007). Final reading outcomes of the national randomized field trial of Success for All. *American Educational Research Journal*, 44(3), 701–731. <https://doi.org/10.3102/0002831207306743>
- Brophy, J., & Good, T. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 328–375). Macmillan.
- Creemers, B. P. M. (1994). *The effective classroom*. Cassell.
- Danielson, C. (2013). *The framework for teaching: Evaluation instrument, 2013 edition*. The Danielson Group LLC.
- Darling-Hammond, L. (2015). Can value added add value to teacher evaluation? *Educational Researcher*, 44(2), 132–137. <https://doi.org/10.3102/0013189X15575346>
- Day, C. (2004). *A passion for teaching*. RoutledgeFalmer.
- Day, C. (2008). Committed for life? Variations in teachers' work, lives and effectiveness. *Journal of Educational Change*, 9(3), 243–260. <https://doi.org/10.1007/s10833-007-9054-6>
- den Brok, P. J. (2001). *Teaching and student outcomes: A study on teachers' thoughts and actions from an interpersonal and a learning activities perspective*. W.C.C.
- den Brok, P., Brekelmans, M., & Wubbels, T. (2004). Interpersonal teacher behaviour and student outcomes. *School Effectiveness and School Improvement*, 15(3–4), 407–442. <https://doi.org/10.1080/09243450512331383262>
- Dobbelaer, M. J. (2019). *The quality and qualities of classroom observation systems* [Doctoral dissertation, University of Twente]. Ipskamp Printing.
- Ellis, E. S., & Worthington, L. A. (1994). *Research synthesis on effective teaching principles and the design of quality tools for educators* (Technical Report No. 5). National Center to Improve the Tools of Educators, University of Oregon.
- Ericsson, K. A. (2006). The influence of experience and deliberate practice on the development of superior expert performance. In K. A. Ericsson, N. Charness, R. R. Hoffman, & P. J. Feltovich (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 685–705). Cambridge University Press.
- Eva, K. W., & Regehr, G. (2005). Self-assessment in the health professions: A reformulation and research agenda. *Academic Medicine: Journal of the Association of American Medical Colleges*, 80(10), S46–S54.
- Faber, J., Luyten, H., & Visscher, A. J. (2017). The effects of a digital formative assessment tool on mathematics achievement and student motivation: Results of a randomized experiment. *Computers & Education*, 106, 83–96. <https://doi.org/10.1016/j.compedu.2016.12.001>
- Ferguson, R. F. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan*, 94(3), 24–28. <https://doi.org/10.1177/003172171209400306>
- Fessler, R., & Christensen, J. C. (1992). *The teacher career cycle: Understanding and guiding the professional development of teachers*. Allyn and Bacon.
- Gess-Newsome, J. (2015). A model of teacher professional knowledge and skill including PCK: Results of the thinking from the PCK Summit. In A. Berry, P. Friedrichsen, & J. Loughran (Eds.), *Re-examining pedagogical content knowledge in science education* (pp. 28–42). Routledge.
- Gitomer, D. H., & Bell, C. A. (2013). Evaluating teachers and teaching. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology: Vol. 3. Testing and assessment in school psychology and education* (pp. 415–444). American Psychological Association. <https://doi.org/10.1037/14049-020>
- Gitomer, D., Bell, C., Qi, Y., McCaffrey, D., Hamre B. K., & Pianta, R. C. (2014). The instructional challenge in improving teaching quality: Lessons from a classroom observation protocol. *Teachers College Record*, 116(6), 1–32.
- Haertel, E. H. (2013). *Reliability and validity of inferences about teachers based on student test scores*. Educational Testing Service Research & Development, Center for Research on Human Capital and Education.
- Hanushek, E. A., & Rivkin, S. G. (2012). The distribution of teacher quality and implications for policy. *Annual Review of Economics*, 4, 131–157. <https://doi.org/10.1146/annurev-economics-080511-111001>

- Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>
- Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, 41(2), 56–64. <https://doi.org/10.3102/0013189X12437203>
- Hu, Y., & van Veen, K. (2020a). Decomposing the observation-based coaching process: The role of coaches in supporting teacher learning. *Teachers and Teaching*, 26(3–4), 280–294. <https://doi.org/10.1080/13540602.2020.1823828>
- Hu, Y., & van Veen, K. (2020b). How features of the implementation process shape the success of an observation-based coaching program: Perspectives of teachers and coaches. *The Elementary School Journal*, 121(2), 283–310. <https://doi.org/10.1086/711070>
- Inspectie van het Onderwijs. (2013). *Professionalisering als gerichte opgave: Verkennend onderzoek naar het leren van leraren* [Deliberate professionalization: An explorative study into teacher learning].
- Kennedy, M. M. (2016). How does professional development improve teaching? *Review of Educational Research*, 86(4), 945–980. <https://doi.org/10.3102/0034654315626800>
- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254–284. <https://doi.org/10.1037/0033-2909.119.2.254>
- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality & Social Psychology*, 77(6), 1121–1134. <https://doi.org/10.1037/0022-3514.77.6.1121>
- Levine, D. U., & Lezotte, L. W. (1995). Effective schools research. In J. A. Banks & C. A. M. Banks (Eds.), *Handbook of research on multicultural education* (pp. 525–547). Macmillan.
- Marzano, R. J., Pickering, D. J., & Pollock, J. E. (2008). *Wat werkt in de klas?* [What works in the classroom?]. Bazalt Educatieve Uitgaven.
- Maulana, R. (2012). *Teacher-student relationships during the first year of secondary education: Exploration of change and link with motivational outcomes in The Netherlands and Indonesia*. [Unpublished doctoral dissertation]. University of Groningen.
- Maulana, R., Opdenakker, M.-C., Stroet, K., & Bosker, R. (2013). Changes in teachers' involvement versus rejection and links with academic motivation during the first year of secondary education: A multilevel growth curve analysis. *Journal of Youth and Adolescence*, 42(9), 1348–1371. <https://doi.org/10.1007/s10964-013-9921-9>
- McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. RAND Corporation.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *The American Council on Education/Macmillan series on higher education. Educational measurement* (pp. 13–103). Macmillan Publishing Co.
- Mintzberg, H. (1979). *The structuring of organizations: A synthesis of the research*. Prentice Hall.
- Muijs, D., & Reynolds, D. (2001). *Effective teaching: Evidence and practice*. SAGE Publications.
- Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, 26(3), 237–257. <https://doi.org/10.3102/01623737026003237>
- Pianta, R. C., & Hamre, B. K. (2009). Classroom processes and positive youth development: Conceptualizing, measuring, and improving the capacity of interactions between teachers and students. *New Directions for Youth Development*, 2009(121), 33–46. <https://doi.org/10.1002/yd.295>
- Purkey, S. C., & Smith, M. S. (1983). Effective schools: A review. *The Elementary School Journal*, 83(4), 427–452. <https://doi.org/10.1086/461325>
- Rosenholtz, S. J. (1989). Workplace conditions that affect teacher quality and commitment: Implications for teacher induction programs. *The Elementary School Journal*, 89(4), 420–439. <https://doi.org/10.1086/461584>
- Rosenshine, B. (2012). Principles of instruction: Research-based strategies that all teachers should know. *American Educator*, 36(1), 12–19.

- Sammons, P., Hillman, J., & Mortimore, P. (1995). *Key characteristics of effective schools: A review of school effectiveness research*. Institute of Education, University of London.
- Scheerens, J. (1992). *Effective schooling: Research theory and practice*. Cassell.
- Scheerens, J. (2016). *Educational effectiveness and ineffectiveness: A critical review of the knowledge base*. Springer. <https://doi.org/10.1007/978-94-017-7459-8>
- Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness*. Elsevier Science.
- Scheerens, J., Seidel, T., Witziers, B., Hendriks, M. A., & Doornekamp, B. G. (2005). *Positioning and validating the supervision framework*. University of Twente.
- Schildkamp, K., & Kuiper, W. (2010). Data-informed curriculum reform: Which data, what purposes, and promoting and hindering factors. *Teaching and Teacher Education*, 26(3), 482–496. <https://doi.org/10.1016/j.tate.2009.06.007>
- Schön, D. A. (1988). From technical rationality to reflection-in-action. In J. Dowie & A. S. Elstein (Eds.), *Professional judgment: A reader in clinical decision making* (pp. 60–77). Cambridge University Press.
- Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16(2), 5–24. <https://doi.org/10.1111/j.1745-3992.1997.tb00585.x>
- Simpson, A. (2018). Princesses are bigger than elephants: Effect size as a category error in evidence-based education. *British Educational Research Journal*, 44(5), 897–913. <https://doi.org/10.1002/berj.3474>
- Strauss, S. (2005). Teaching as a natural cognitive ability: Implications for classroom practice and teacher education. In D. B. Pillemer & S. H. White (Eds.), *Developmental psychology and social change: Research, history and policy* (pp. 368–388). Cambridge University Press.
- Timmermans, A. C. (2012). *Value added in educational accountability: Possible, fair and useful?* [Doctoral dissertation, University of Groningen]. GION, Gronings Instituut voor Onderzoek van Onderwijs, Rijksuniversiteit Groningen.
- Trochim, W. M. K. (2006). *Introduction to validity*. <http://www.socialresearchmethods.net/kb/introval.php>
- Troman, G., & Woods, P. (2001). *Primary teachers' stress*. RoutledgeFalmer.
- Tschannen-Moran, M., Woolfolk Hoy, A., & Hoy, W. K. (1998). Teacher efficacy: Its meaning and measure. *Review of Educational Research*, 68(2), 202–248. <https://doi.org/10.3102/00346543068002202>
- van der Lans, R. M., van de Grift, W. J. C. M., van Veen, K., & Fokkens-Bruinsma, M. (2016). Once is not enough: Establishing reliability criteria for feedback and evaluation decisions based on classroom observations. *Studies in Educational Evaluation*, 50, 88–95. <https://doi.org/10.1016/j.stueduc.2016.08.001>
- van der Lans, R. M., van de Grift, W. J. C. M., & van Veen, K. (2015). Developing a teacher evaluation instrument to provide formative feedback using student ratings of teaching acts. *Educational Measurement: Issues and Practice*, 34(3), 18–27. <https://doi.org/10.1111/emip.12078>
- van der Scheer, E. A., Bijlsma, H. J. E., & Glas, C. A. W. (2019). Validity and reliability of student perceptions of teaching quality in primary education. *School Effectiveness and School Improvement*, 30(1), 30–50. <https://doi.org/10.1080/09243453.2018.1539015>
- van Geel, M., Keuning, T., Visscher, A. J., & Fox, J.-P. (2016). Assessing the effects of a school-wide data-based decision-making intervention on student achievement growth in primary schools. *American Educational Research Journal*, 53(2), 360–394. <https://doi.org/10.3102/0002831216637346>
- van Merriënboer, J. J. G., & Kirschner, P. A. (2007). *Ten steps to complex learning: A systematic approach to four-component instructional design*. Lawrence Erlbaum Associates.
- van Tartwijk, J., Brekelmans, M., Wubbels, T., Fisher, D. L., & Fraser, B. J. (1998). Students' perceptions of teacher interpersonal style: The front of the classroom as the teacher's stage. *Teaching and Teacher Education*, 14(6), 607–617. [https://doi.org/10.1016/S0742-051X\(98\)00011-0](https://doi.org/10.1016/S0742-051X(98)00011-0)
- Visscher, A. J. (2015). *Over de zin van opbrengstgericht werken in het onderwijs* [On the value of data-based decision-making in education]. GION onderzoek/onderwijs.

- Visscher, A. J. (2021). On the value of data-based decision making in education: The evidence from six intervention studies. *Studies in Educational Evaluation*, 69, Article 100899. <https://doi.org/10.1016/j.stueduc.2020.100899>
- Visscher, A. J., & Coe, R. (2003). School performance feedback systems: Conceptualisation, analysis, and reflection. *School Effectiveness and School Improvement*, 14(3), 321–349. <https://doi.org/10.1076/sesi.14.3.321.15842>
- Walberg, H. J., & Haertel, G. D. (1992). Educational psychology's first century. *Journal of Educational Psychology*, 84(1), 6–19. <https://doi.org/10.1037/0022-0663.84.1.6>
- Ward, J. R., & McCotter, S. S. (2004). Reflection as a visible outcome for preservice teachers. *Teaching and Teacher Education*, 20(3), 243–257. <https://doi.org/10.1016/j.tate.2004.02.004>
- Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The Widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness: Executive summary, second edition*. New Teacher Project.
- Weiss, C. H. (1998). Improving the use of evaluations: Whose job is it anyway? In A. J. Reynolds & H. J. Walberg (Eds.), *Advances in educational productivity* (Vol. 7, pp. 263–276). JAI Press.
- William, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, 37(1), 3–14. <https://doi.org/10.1016/j.stueduc.2011.03.001>