Validity and reliability of student perceptions of teaching quality in primary education

van der Scheer, Emmelien A.; Bijlsma, Hannah J. E.; Glas, Cees A. W.

Published in:
School Effectiveness and School Improvement

DOI:
10.1080/09243453.2018.1539015

Publication date:
2019

Routledge
Taylor & Francis Group

School Effectiveness and School Improvement
An International Journal of Research, Policy and Practice

ISSN: 0924-3453 (Print) 1744-5124 (Online) Journal homepage: https://www.tandfonline.com/loi/nses20

Validity and reliability of student perceptions of teaching quality in primary education

Emmelien A. van der Scheer, Hannah J. E. Bijlsma & Cees A. W. Glas

To cite this article: Emmelien A. van der Scheer, Hannah J. E. Bijlsma & Cees A. W. Glas (2019) Validity and reliability of student perceptions of teaching quality in primary education, School Effectiveness and School Improvement, 30:1, 30-50, DOI: 10.1080/09243453.2018.1539015

To link to this article: https://doi.org/10.1080/09243453.2018.1539015

Published online: 19 Nov 2018.

SCHOOL EFFECTIVENESS AND SCHOOL IMPROVEMENT
2019, VOL. 30, NO. 1, 30–50
https://doi.org/10.1080/09243453.2018.1539015

Routledge
Taylor & Francis Group

ARTICLE

OPEN ACCESS Check for updates

Validity and reliability of student perceptions of teaching quality in primary education

Emmelien A. van der Scheer [a], Hannah J. E. Bijlsma [b] and Cees A. W. Glas[c]

[a]GION education/research, University of Groningen, Groningen, The Netherlands; [b]Department of Teacher Development, University of Twente, Enschede, The Netherlands; [c]Department of Research Methodology, Measurement and Data Analysis, University of Twente, Enschede, The Netherlands

ABSTRACT

A Bayesian IRT-model approach was used to investigate the validity and reliability of student perceptions of teaching quality. Furthermore, the student perceptions were compared with ratings of teaching quality by external observers. Grade 4 students ($n = 675$) filled out a questionnaire that was used to measure their opinions about the lessons of their teachers. Three lessons of 39 teachers were recorded and rated by 4 raters. The analyses showed that student perception and lesson observation scales fit best in an 11-dimensional model, which was an indication of construct validity and discriminant validity. Student perception scales were reliable, although not all items contributed to the scales to the same extent. Student ratings and lesson observations scores generally correlated moderately (ranging from $r = .18$ to $r = .50$). Higher correlations were found for scales with a similar content; however, no clear pattern was apparent. Suggestions for future research are presented.

KEYWORDS

Student perceptions; classroom observations; teaching quality

Introduction

Determining teaching quality reliably and validly serves important improvement and accountability purposes (Timperley, Wilson, Barrar, & Fung, 2007). In primary education, teaching quality is predominantly measured using classroom observations (Goe, Bell, & Little, 2008). However, classroom observations are time consuming as obtaining reliable measurements requires multiple lesson observations carried out by multiple trained observers (e.g., Hill, Charalambous, & Kraft, 2012). As this is generally not feasible for school leaders, school inspectors, and researchers, it is common for only one lesson to be observed, which causes problems regarding the reliability (and validity) of the measures.

A less time-consuming method is the use of student perceptions of teaching quality (Goe et al., 2008). Students experience their teacher daily and thus may be an important source of information regarding the teaching qualities of their teachers (Ferguson & Danielson, 2014; Gaertner, 2014; Peterson, Wahlquist, & Bone, 2000). Furthermore, students have experienced various teachers, and as such can base their judgements on comparative observations (Goe et al., 2008).

CONTACT Emmelien A. van der Scheer e.a.van.der.scheer@rug.nl GION education/research, University of Groningen, Groningen, The Netherlands

However, student perceptions are rarely used in primary education for measuring teaching quality (Hamre & Pianta, 2010; Kyriakides, 2005). Although recent studies have shown that student perceptions of teaching quality can provide reliable and valid information for both formative evaluation and research purposes (Burniske & Meibaum, 2012; Ferguson & Danielson, 2014; Kane, McCaffrey, Miller, & Staiger, 2013; Kyriakides, 2005; Peterson et al., 2000), there are concerns about the validity and reliability of the perceptions of younger children (De Jong & Westerhof, 2001; Ferguson, 2012, Kunter & Baumert, 2006).

These concerns generally involve the discriminant validity of student perceptions, that is, the extent to which students are able to discriminate between the different facets of teaching (Fauth, Decristan, Rieser, Klieme, & Büttner, 2014). Because studies into student perceptions of teaching quality in primary education are limited, additional research is required to identify the extent to which such perceptions could be used for evaluating teaching quality. More insight into the validity and reliability of student perceptions of teaching quality in primary education, and how these perceptions relate to external observer ratings, is especially important (Maulana & Helms-Lorenz, 2016; Van der Lans, 2017). Therefore, the following research questions are addressed in this study:

(1) To what extent do fourth-grade students' perceptions of teaching quality show construct and discriminant validity?
(2) To what extent are fourth-grade students' perceptions of teaching quality reliable at the scale and item levels?
(3) How consistent are student perceptions of teaching quality with observer ratings of lessons?

Theoretical framework

First, some concerns about the validity and reliability of student perceptions of teaching quality in primary education are described. This is followed by a discussion of what is known about the extent to which the results of student perceptions and observations match. To indicate whether student perceptions of teaching quality can be used as measures of teaching quality, it is important to know which aspects of teaching are relevant to focus on. Therefore, the characteristics of effective lessons that are generally measured using observation schemes are also presented here.

Student perceptions of teaching quality in primary education

As was already mentioned in the introduction of this article, students may be a valuable source of information regarding teachers' teaching qualities because their views are based on many lessons (De Jong & Westerhof, 2001). Moreover, student perceptions are cost efficient as multiple raters (e.g., 20–30 students) all provide their opinion at one moment in time, which also reduces rater bias (De Jong & Westerhof, 2001; Goe et al., 2008; Kyriakides, 2005).

Peterson et al. (2000) found that student perception questionnaires used at various levels of the education system (primary school, middle school, and high school) were reliable and valid teacher evaluation measures. Although several studies into student

perceptions have also provided evidence that students are generally capable of discriminating between teaching quality constructs (Fauth et al., 2014; Greenwald, 1997; Kyriakides, 2005; Wagner, Göllner, Helmke, Trautwein, & Lüdtke, 2013), two aspects of discriminant validity are still debated with respect to student perceptions in primary education. First, whether students are able to distinguish between different teaching quality constructs and, second, whether student perceptions are confounded with teacher popularity (Fauth et al., 2014). Related to the first concern, younger students are untrained at rating teaching behavior, thus they may have a less sophisticated understanding of different teaching aspects compared to a trained observer (Van der Lans, Van de Grift, & Van Veen, 2015). Therefore, the extent to which they can discriminate between different teaching constructs may be limited (Fauth et al., 2014). The concern regarding teacher popularity is that popular teachers may receive higher student perception scores regardless of their teaching quality (Ben-Chaim & Zoller, 2001). Combined with the potentially limited ability to discriminate between different teaching quality constructs, student perceptions may simply be a popularity contest rather than a measure of teaching quality (Fauth et al., 2014).

Combining multiple measurements for establishing validity and reliability

It is to be expected that students and external observers have differing views concerning a teacher's teaching quality because of differences in background, knowledge, and aims (Patton, 1980). External observers generally do not have a(n) (emotional) connection with the observed teacher and may have been trained and familiarized with the use of a predetermined observation scheme. Furthermore, their aim is generally to assess teaching quality as objectively as possible during one or more lessons (Kane & Staiger, 2012). If trained well, observers have knowledge regarding what effective teaching looks like, and thus can assess teaching behavior from an evidence-based perspective.

On the other hand, student perceptions of teaching quality reflect the perception of the target group, resulting in data based on subjective frames of references (Kane & Staiger, 2012). Students have less knowledge of effective teaching than a trained observer does, and they therefore provide information about their experience of the teaching activities of their teachers. Moreover, students normally have an emotional connection with their teachers, which might lead to socially desirable answers from students (Maulana & Helms-Lorenz, 2016).

However, research into the extent to which both measurements relate is limited, and results are often subject to critique. A study by De Jong and Westerhof (2001) showed that student ratings of teaching quality were as reliable as ratings from external observers. They however found that both measurements correlated slightly (ranging from $r = -.17$ to $r = .38$ with only 4 out of 30 correlations at or above $r = .30$). The authors explained this finding by indicating that the content of the scales differed substantially, thus the scales used across the instruments were incomparable. Van der Lans (2017) also found a low correlation between classroom observations and student questionnaires ($r = .26$ overall; $r = .34$ after removal of misfitting items), consistent with other studies in which student perceptions and observer ratings of teaching quality were statistically related to each other (Howard, Conway, & Maxwell, as cited in Van der Lans, 2017; Maulana & Helmz-Lorenz, 2016). However, in these studies the researchers observed only one lesson and/or worked with

only one rater (instead of several observations and raters, which are required for the reliable measurement of teaching quality; Hill et al., 2012).

Therefore, the extent to which observer ratings and student perceptions in primary education are consistent with each other is still unclear, especially if similar teaching dimensions are used and if several lesson observations per teacher are rated by multiple observers. In the current study, the results of multiple lesson observations per teacher (assessed by multiple raters) were statistically related to student perceptions of teaching qualities in primary education. To clarify whether student perceptions of teaching quality can be used as a measure for teacher evaluation, similar teaching dimensions were measured by means of student perceptions and classroom observations. To indicate which teaching constructs are relevant to focus on in evaluations of teaching quality, the characteristics of effective lessons are described below.

Characteristics of effective lessons

The study of effective classroom practices and "what works" in the classroom has been central to the measurement of teaching quality (Reynolds et al., 2014). Several studies have been conducted to identify effective teaching practices (e.g., Creemers, 1994; Fauth et al., 2014; Hattie, 2008; Muijs et al., 2014; Pianta & Hamre, 2009; Reynolds et al., 2014; Sammons, Hillman, & Mortimore, 1995; Van de Grift, 2007), and a variety of teacher behaviors which promote student learning were identified. These can be categorized into three basic teaching dimensions (Day et al., 2008).

First, providing a supportive, positive, and inclusive classroom climate is important (Fraser, 1998) and promoting positive teacher–pupil relationships through praise and feedback (Den Brok, Brekelmans, Levy, & Wubbels, 2002).

Second, classroom management quality can be used to identify an effective teacher (Van de Grift, 2007). Classroom management is about having clear classroom rules and routines, preventing disruptive behavior, and about having well-organized and structured lessons (Day et al., 2008).

Third, the teacher's instructional approach is important. This involves connecting the lesson to what students already know, explaining the subject matter in such a way that students understand it, engaging students by means of assignments and activities, and providing feedback (Hattie & Timperley, 2007; Hollingsworth & Ybarra, 2009; Rosenshine, 1995).

Van de Grift (2007) adds two teaching dimensions to these three features of effective lessons. According to Van de Grift, effective teachers also teach learning strategies to their students. They support the higher level thinking and metacognitive learning of their students (Arends, 2009). Teaching then is broken down into small steps wherein teachers provide simplified problems, model problem-solving strategies, and think aloud when solving a problem (Rosenshine, 1995). Moreover, Van de Grift (2007) and Maulana, Helms-Lorenz, and Van de Grift (2014) stress the importance of adaptive instruction in the classroom: teachers adapting their teaching activities to the (varying) needs of their students. Another feature of effective lessons is ensuring that lessons are meaningful to students (Keuning & Van Geel, 2016; Kyriakides, Campbell, & Gagatsis, 2000). This means that teachers set clear lesson goals and clarify them to the students at the beginning of the lesson (*what* they are learning and *why*), that they make sure that all lesson activities relate to these goals, and that the goals are evaluated (Locke & Latham, 2002).

In summary, several important dimensions of teaching quality were discussed: providing a positive and inclusive classroom climate, quality of classroom management, clear and activating instructional approach (the three basic dimensions), adaptive instruction, teaching learning strategies, and goal orientation.

Method

Design

The data used in this study were collected as part of an intervention study. The teachers participated in a data-based decision-making intervention during the school year 2013–2014 (e.g., Van der Scheer & Visscher, 2018). The lessons of participating teachers were recorded prior to the intervention, either at the end of school year 2012–2013 or at the start of school year 2013–2014. The students completed a student perception questionnaire on the teaching quality of their teachers after the first few weeks of the school year 2013–2014.

Participants

In this study, 31 teachers (64.5% women) from 27 primary schools participated. Eight teachers were part of a teacher pair (two teachers teaching the same class). The teachers had, on average, 13.29 years (SD = 10.06) of teaching experience. Each teacher taught a fourth-grade class or a multigrade class including fourth-grade students.

Their 675 students (52.4% boy) filled out the student perception questionnaire. The majority of students were fourth-grade students (83.1%), the other students were either in Grade 3 or in Grade 5. The average class size was 19.35 students (SD = 4.98).

Instruments

Observation scheme

The validated ICALT (International Comparative Analysis of Learning and Instruction) lesson observation instrument was used by the observers to rate the teachers' teaching quality (Van de Grift, 2007). In line with the study of Van der Scheer, Glas, and Visscher (2017), the following six scales were used: safe and stimulating learning climate, efficient classroom management, clear and activating instruction, adaption of instruction, teaching learning strategies, and engagement of students. The first five constructs are discussed in the theoretical framework of this study as effective teaching dimensions. The latter is about how lesson observers rated students' behavior (their engagement in the lesson) instead of teacher behavior, as student engagement is an important prerequisite for learning (Van de Grift, 2007). All 35 items were scored on a 4-point Likert scale (ranging from *predominantly weak* to *predominantly strong*). The content of the ICALT instrument is discussed extensively in other studies (see, e.g., Maulana et al., 2014; Van de Grift, 2007, 2014; Van de Grift, Van der Wal, & Torenbeek, 2011).

The student perception questionnaire

The student perception questionnaire used in this study was originally developed to evaluate the effects of a data-based decision-making (DBDM) intervention (Keuning & Van Geel, 2016; Van der Scheer, 2016). The development of the questionnaire was inspired by the Tripod 7Cs instrument, used in the large-scale Measures of Effective Teaching project (Ferguson & Danielson, 2014). The translated Tripod 7Cs questionnaire was then piloted with 59 primary school teachers, and the results were discussed with an expert group of primary education teachers. Based on the gathered feedback, the instrument was further adapted to the Dutch context (in terms of the items used and the wording of the items). The resulting questionnaire consists of 36 items, scored on a 5-point Likert scale (ranging from *no never* to *yes, always*) and includes the following five scales: classroom climate (6 items), classroom management (9 items), instruction (10 items), goal orientation (6 items), and challenging students (5 items). The first four are discussed in the theoretical framework of this study as effective teaching dimensions. The fifth scale measures the extent to which students feel challenged by their teacher as a result of the expectations teachers have of them. Teacher expectations are identified as one of the most important factors for student learning in educational effectiveness research (Muijs et al., 2014).

All 11 scales in the lesson observation instrument and the student perception questionnaire were included in this study. An overview of the observer scales and the student perception scales is provided in Table 1. It shows that the content of the first three scales is comparable for both instruments. As discussed in the theoretical framework, these three scales represent the three general dimensions of effective teaching. The content of the remaining teaching quality scales differs between the two instruments due to the different backgrounds of the questionnaires. However, the observers' scale regarding the adaptation of instruction also included items that are related to goal orientation.

Data collection

Each teacher was recorded during three mathematics lessons, and the recordings were rated by four observers. Although both multiple raters and multiple lessons are necessary to obtain reliable estimates of teachers' teaching skills (e.g., Hill et al., 2012), these requirements are seldom met. The lessons were recorded using the IRIS Connect video system. This system consists of two iPods; one iPod recorded what the teacher did, while the other iPod simultaneously recorded the students. The recordings were uploaded to an online environment, where both observers and teachers could view the recordings.

Table 1. Overview of scales in observation scheme and student perception questionnaire.

Scale	ICALT observation scheme (35 items)	Student perceptions (36 items)
1	Classroom climate	
2	Classroom management	
3	Instruction	
4	Adaption of instruction	Challenging students
5	Learning strategies	Goal orientation
6	Engagement of students	

To avoid order bias, each observer was provided with a list of the lessons to be rated in a specific sequence; the sequence of the recordings had been ordered randomly (Shadish, Cook, & Campbell, 2002). Each lesson was rated by each observer; however, one observer could not assess two recordings.

The three observers were trained for 3 days in the use of the ICALT observation instrument. As part of the training, observers viewed six recorded lessons independently. The results were discussed to achieve consensus on the use of the scale. A fourth experienced observer was added later. Observer variance accounted for a low percentage of the total variance, which indicates a high reliability of the estimates of the teachers' teaching quality (Van der Scheer et al., 2017).

Students filled out the student perception questionnaire about the teaching quality of their teacher. To avoid test scores being influenced by the presence of their teacher, it was recommended that another teacher substituted the teacher during the administration of the questionnaire. Afterwards, teachers received a report including their average scores given by their students and a comparison of their scores with the scores of other teachers.

Analyses

Nowadays, latent variable models in general and item response theory (IRT; see, e.g., Lord, 1980) in particular have become the standard statistical tools in educational measurement. For instance, all large-scale international educational surveys, such as the Programme for International Student Assessment (PISA), the Trends in International Mathematics and Science Study (TIMSS), the Progress in International Reading Literacy Study (PIRLS), the International Computer and Information Literacy Study (ICILS), and the Programme for the International Assessment of Adult Competencies (PIAAC), use this methodology (see, e.g., Rutkowski, von Davier, & Rutkowski, 2014). This methodology will also be used for the analyses in the current research.

IRT is used to model the responses of the students to the questionnaire and observers to the observation scheme. The advantage of using an IRT measurement model over a simple sum-score model is that it explicitly accounts for the discrete nature of the item responses. Furthermore, the parameter estimates give insight into how the individual items contribute to the overall reliability. Each of the 11 scales (see Table 1) was associated with a unidimensional latent proficiency variable (the estimate of a teacher's teaching quality per scale), that is, with an IRT model. Below, the designations, proficiency, and teaching quality will be used interchangeably.

To assess the extent to which student perceptions of teaching quality show construct and discriminant validity and answer the first research question, the five latent variables related to the five subscales used by the students were correlated. This is completely analogous to performing a confirmatory factor analysis on categorical data. The correlations between the scales should be lower than $r = .85$ to indicate discriminant validity (Brown, 2015). To answer the second research question, about the reliability of student perceptions of teaching quality, the global (on scale level) and local (on item level) reliability of the student perception scales was estimated. For the third research question, the correlations between the teacher proficiencies for each scale as rated by external observers and the student perceptions of teaching quality were estimated.

Since the structure of the data is different for the student perceptions (a multilevel structure, with multiple students providing their opinions about the teachers' teaching quality) and the observer ratings (three lessons per teacher, rated by four raters), two different underlying models of teachers' proficiency had to be defined. Below, first a description of *the general model* is provided, followed by a description of the *student perceptions model* and the *observer ratings model*.

Finally, the complete model, that is, the IRT models for the observation scales and their covariance structure, was estimated in a Bayesian framework using OpenBugs. The motive for this choice rather than traditional standard software for latent variable modeling is that OpenBugs allows the users to completely specify their own model, taking into account all dependencies.

The general model

The relation between the ratings by the external observers and the ratings by the students was modeled as follows. The observation instrument used by the students consisted of $K^{(s)} = 36$ items distributed across five subscales, while the observation instrument used by the observers consisted of $K^{(o)} = 35$ *items* distributed across six subscales (see Table 1). The items (indexed by the subscript i) were scored polytomously (more than two response categories). The response categories were indexed as $j = 0, 1, 2, \ldots, M$, where M was the highest category. For the student questionnaire, $M = M^{(s)} = 4$; for the observers, $M = M^{(o)} = 3$.

The item responses were assumed to load on latent variables, one for every subscale, with a total of $Q = 11$ subscales, 5 for the students and 6 for the observers. So, the proficiency of every teacher (indexed n) is modeled by a Q-dimensional vector θn = θn1, ... , θnq, ... ,θnQ. These proficiency vectors have a normal distribution with expectation zero and a covariance matrix Σ, that is,

$$\theta_n \sim N(0, \Sigma) \tag{1}$$

The covariance matrix reflects the relation between various subscales and is the primary concern of the analyses. In the Results section, the covariance matrix is converted to a correlation matrix. It is an indicator for the discriminant validity of the student perceptions (correlation between the students' dimensions in the matrix) and the consistency between the observer ratings and student perceptions (correlation between the students' and observers' dimensions).

An IRT model was used to transform the discrete item responses X_i ($X_i = j, j = 0, \ldots, M$) to continuous latent values θ. The specific IRT model used was the generalized partial credit model (GPCM; Muraki, 1992). In the GPCM, the probability of observing a score X_i in a response category j ($j = 0, \ldots, M$) is defined as:

$$P(X_i = j|\theta) = \Psi_{ij}(\theta) = \frac{\exp(ja_i\theta - b_{ij})}{1 + \sum_{h=1}^{M} \exp(ha_i\theta - b_{ih})} \tag{2}$$

The item parameters b_{ij} ($j = 0, \ldots, M$, $b_{i0} = 0$ to identify the model) are related to the location of the items on the latent θ scale. The parameter a_i is called a discrimination parameter and reflects the extent to which the item depends on the latent variable θ.

The multidimensional IRT model used here is closely related to a confirmatory factor analysis model (Takane & De Leeuw, 1987). Therefore, translated to factor-analytic terminology, the discrimination parameter is a factor loading.

The model for the student perceptions

The model for the student perceptions is a multilevel model with two levels. The first level is an IRT model for a response regarding teacher n, assessed by student s ($s = 1, \ldots, N_n$ where N_n is the number of students in the class of teacher n) on scale q, say the response X_{nsqi}. The model is a special version of the general model defined by Formula (2). Therefore, the probability of a response on item i in category j is defined by $P(X_{nsqi} = j|\theta_{nsq}) = \Psi_{ij}(\theta_{nsq})$, where the function $\Psi_{ij}(.)$ is defined as in Formula (2). The second-level model entails that the variables θ_{nsq} have a normal distribution, that is,

$$\theta_{nsq} \sim N(\theta_{nq}, \sigma^2_{0q}) \tag{3}$$

So, the mean is equal to the teacher's proficiency θ_{nq} on scale q, and the variance σ^2_{0q} indicates the degree to which students differ in their assessment on this scale. Notice that this is the definition of a multilevel model: Students are nested under a teacher and are assumed random effects.

The model for the observers' ratings

The model for the observer was built up along the same lines, but here three different time points ($t = 1, \ldots, T, T = 3$), and four observers ($r = 1, \ldots, R, R = 4$) had to be accounted for. As such, the responses were modeled through a combination of an IRT model and a generalizability theory (GT) model (Brennan, 1992; Brennan & Johnson, 1995), defined on a latent variable (Glas, 2012). Let X_{nqrti} be the response pertaining to teacher n by observer r on time point t, on item i of scale q. Then the response model is defined by $P(X_{nqrti} = j|\theta_{nqrt}) = \Psi_{ij}(\theta_{nqrt})$, where, again, the function $\Psi_{ij}(.)$ is defined as in Formula (2). The GT model is defined as

$$\theta_{nqrt} = \theta_{nq} + \tau_{1qr} + \tau_{2qt} + \tau_{3nqr} + \tau_{4nqt} + \tau_{5qrt} + \varepsilon_{nqrt},$$

where all components with their interpretation and distribution are summarized in Table 2.

For each of the scales, the parameters of the combined IRT model, the multilevel model, and the GT model were concurrently estimated in a Bayesian framework using OpenBugs (Version 3.2.3., rev 2012).

Table 2. Generalizability theory model: parameters, interpretation, and distributions.

Parameter	Interpretation	Distribution
θ_{nq}	Proficiency level of teacher n on scale q	Formula (1)
τ_{1qr}	Main effect observer	$N(0, \sigma^2_{1q})$
τ_{2qt}	Main effect time point	$N(0, \sigma^2_{2q})$
τ_{3nqr}	Interaction effect teacher and observer	$N(0, \sigma^2_{3q})$
τ_{4nqt}	Interaction effect teacher and time point	$N(0, \sigma^2_{4q})$
τ_{5qrt}	Interaction effect observer and time point	$N(0, \sigma^2_{5q})$
ε_{nqrt}	Residual, confounded with three-way interaction teacher, observer, and time point	$N(0, \sigma^2_{\varepsilon q})$

Model comparisons

In this section, the fit of the model is evaluated. A more complex model fits data better than a simpler model; however, at some point adding more complexity destroys the interpretability and reliability of the analyses. Therefore, we apply a methodology for the evaluation of model fit that penalizes overly complex models. The model used above for the analyses was a between-items 11-dimensional model, as six observer scales and five student perception scales were defined. It is between-item multidimensional because the response probabilities as defined by Formula (2) depend on only one latent variable. For instance, $\Psi_{ij}(\theta_{nqrt})$ depends on one specific value θ_{nqrt} for a unique teacher n, on one scale indexed q, by observer r at time point t. To obtain an impression of the fit of the model, the model was compared to three alternative models. The first two were simpler than the between-items 11-dimensional model and had fewer parameters. In these models, the student and observer sub-scales (as presented in Table 1) were not taken into account to investigate both construct and discriminant validity. The first one was a unidimensional model where all 11 dimensions were collapsed into a single dimension. The hypothesis here was that all items measured the same dimension for both the students and the observers. The second model assumed two correlated dimensions, wherein it was hypothesized that two dimensions could be distinguished: one for the student perceptions and one for the external observer ratings.

The models were compared using the so-called deviance information criterion (DIC). The DIC is the sum of two components, the expected deviance denoted by $\bar{D}$ and a penalty for the number of parameters in the model, denoted by p_D. As is done in the more common likelihood-based approach, a lower deviance is an indication of better model fit. Further, p_D favors models with a smaller number of parameters. The third model, with which the between-items 11-dimensional model was compared, was a slightly more general model, which was called the 11-dimensional bi-factor model. It still has 11 dimensions. However, every item now loads on two dimensions, one general teaching quality dimension, say θ_0, and one scale-specific dimension, say θ_q. The model is defined as:

$$P(X_i = j|\theta_0, \theta_q) = \Psi_{ij}(\theta_0, \theta_q) = \frac{\exp\left(j\left(a_{i0}\theta_0 + a_{i1}\theta_q\right) - b_{ij}\right)}{1 + \sum_{h=1}^{M} \exp\left(j\left(a_{i0}\theta_0 + a_{i1}\theta_q\right) - b_{ih}\right)}$$

Note that compared to the between-items 11-dimensional model, there are more discrimination parameters in the model. However, both for the students and observers scales, the items pertaining to one of the subscales must be unidimensional to identify the model. The results are given in Table 3.

Table 3. Model comparisons using the DIC.

Model	$\bar{D}$	P_D	*DIC*
Unidimensional	68,560	2156	70,716
Two-dimensional	67,973	2161	70,134
11-dimensional	66,190	3104	69,294
11-dimensional bi-factor	66,160	3146	69,400

Although the DIC values, as presented in Table 3, are quite close, the between-items 11-dimensional model has the lowest DIC value; so, this is the preferred model. Thus, the confirmatory factor analyses showed that the model in which the six observer scales and five student perception scales were distinguished best fitted the data, and this model was therefore used for the analyses.

It will now be described how the scale reliability (of both the observer and students scales) and the local reliability (of the items of students scales) were estimated to answer the second research question regarding the reliability of student perceptions.

Reliability of the observer scale

The reliability of the external observer scales was high (ranging from 0.82, with a posterior variance of 0.05, to 0.95, with a posterior variance of 0.01). For the external observer ratings, the reliability is equal to the ratio of the systematic variance, that is, the variance of the teachers and the systematic variance plus all error variance components (for more information on reliability indices in generalizability theory, see Brennan, 1992).

Reliability of the student scale

For the student ratings, the reliability is equal to:

$$\rho_q = \frac{\Sigma_{qq}}{\Sigma_{qq} + \frac{\sigma^2_{0q}}{\bar{N}_s}}$$

where Σ_{qq} is the variance of the teachers on scale q, that is, the q-th diagonal element of the covariance matrix Σ defined in Formula (2), and σ^2_{0q} is the variance of the student ratings as defined in Formula (3). Further, $\bar{N}_s$ stands for the average number of students in a class.

Not only the reliability for *the entire scale* (global reliability) was estimated but also the local reliability. The local reliability is indicated by the contributions of *individual items to the reliability*, which depends on both the discrimination and item location parameters. High discrimination and a location close to a specific θ lead to high reliability at that point. Local reliability is expressed in "Information". The information that an item i attributes to the reliability with which θ is estimated, is equal to:

$$\text{Info}_i(\theta) = a_i^2 \sum_{j=1}^{M} \left\{ j\Psi_{ij}(\theta) \left[j - \sum_{h=1}^{m} h\Psi_{ih}(\theta) \right] \right\}$$

As will be discussed in the Results section, Table 6 presents the expected item information computed over the posterior distribution of θ.

Results

The discriminant validity of student perceptions of teaching quality is described first to answer the first research question. Next, the global and local reliabilities (second research question) of the student perception scales are presented. To answer the third research question, the results of the analyses of the consistency (correlations) between the observers' ratings of teaching quality and the student perceptions of teaching quality are presented.

Discriminant validity of student perceptions

Table 4 gives the correlations between the student perceptions scales (see Formula [1]). The table shows that the scales correlate moderately to strongly (ranging from $r = .42$ to $r = .74$). The correlation between the classroom climate scale and the instruction scale is highest ($r = .74$). Apparently, classes in which students reported that they experienced a safe classroom climate also felt that their teacher's instruction was clear. A moderate correlation was found between goal orientation and instruction ($r = .65$), two aspects of teaching that relate to the same lesson phase. The lowest correlation was found between the classroom management scale and the goal-orientation scale ($r = .42$). This indicates that students' experience of classroom management correlated limitedly to the extent to which their teachers communicated the lesson goals with them.

Table 4. The correlations between the student perception scales.

	S1 – Climate	S2 – Management	S3 – Instruction	S4 – Challenging students	S5 – Goal orientation
S1 – Classroom climate	1.00				
S2 – Classroom management	.63	1.00			
S3 – Instruction	.74	.59	1.00		
S4 – Challenging students	.62	.52	.59	1.00	
S5 – Goal orientation	.56	.42	.65	.53	1.00

Global reliability student perceptions – scale level

Table 5 presents the estimates of the reliabilities of the student perception scales and shows that the global reliability of each of the student perception scales is sufficient, as the reliabilities ranged from .80 to .91.

Table 5. The reliabilities of the student perception scales.

	Reliability (post *SD*)
S1 – Classroom climate	.91 (0.02)
S2 – Classroom management	.90 (0.03)
S3 – Instruction	.88 (0.03)
S4 – Challenging students	.80 (0.06)
S5 – Goal orientation	.91 (0.02)

Local reliability student perceptions – item level

Table 6 shows the estimates of the student scales' item parameters. The columns labeled d and a respectively refer to the item location and discrimination parameters (as described in the *general model* description). For items with a higher location parameter (d), it is harder to

receive a high score. For Item 5 ("My teacher seems to know if something is bothering me"), it is relatively hard for teachers to receive a positive judgement from their students. On the other hand, students generally report that they know when they can ask their teacher questions during work time (Item 11) as this item has a low location parameter.

Items with high discriminations (a) load high on their relevant scale. So, Item 2 ("I like the way my teacher treats me when I need help") has a high loading on Student scale 1 (classroom climate), whereas Item 6 ("I like this class") has a low loading on that scale.

Table 6. Estimates of item parameters of the student perception scales.

		d		a		Information	
Item		Est.	*SD*	Est.	*SD*	Est	*SD*
Scale 1 – Classroom climate							
1	My teacher is nice to me when I ask questions	0.31	0.04	1.52	0.29	594.8	227.7
2	I like the way my teacher treats me when I need help	1.12	0.03	1.86	0.35	861.9	282.5
3	My teacher wants me to do well at school	−1.88	0.11	1.03	0.22	192.2	77.4
4	When I am sad or angry, my teacher helps me so I will feel better	2.87	0.08	0.94	0.18	474.9	163.2
5	My teacher seems to know if something is bothering me	4.49	0.21	0.66	0.13	274.1	105.8
6	I like this class	−0.18	0.22	0.40	0.08	115.6	44.1
Scale 2 – Classroom management							
7	We start the lessons on time	−1.06	0.20	0.63	0.15	215.9	86.5
8	When my teacher explains something, it takes a long time before everybody is listening	5.91	0.35	0.59	0.14	201.5	83.2
9	When we are working individually, it is quiet in the classroom	2.28	0.10	1.18	0.27	654.1	232.8
10	We have clear rules in the classroom	−3.98	0.15	0.71	0.18	199.7	84.3
11	I know when I can ask my teacher questions during work time	−5.25	0.22	0.47	0.12	125.8	54.6
12	Our classroom is neat and tidy	0.68	0.13	0.89	0.20	380.9	140.0
13	Everybody pays attention when my teacher explains something	2.83	0.08	1.64	0.37	940.8	309.5
14	Everybody in our class works hard	1.02	0.08	1.27	0.29	680.7	222.8
15	Other classmates disturb me when we work individually	3.60	0.45	0.38	0.10	124.8	55.2
Scale 3 – Clear instruction							
16	My teacher explains difficult things clearly	−0.82	0.11	0.57	0.13	195.8	69.0
17	If I don't understand something, my teacher explains it another way	1.52	0.12	0.73	0.15	307.7	108.5
18	When my teacher explains something, I get it right away	0.47	0.43	0.48	0.11	105.9	43.1
19	My teacher wants me to explain how I got to my answer	1.21	0.14	0.57	0.12	228.3	84.7
20	My teacher knows how he/she can best explain something to me	3.72	0.13	0.81	0.16	427.2	144.8
21	My teacher knows when I understand something, and when I do not	0.74	0.05	1.48	0.29	697.1	223.7
22	My teacher helps me if I do not understand something	−0.50	0.03	1.48	0.30	609.7	197.8
23	My teacher asks questions to be sure I understand	1.96	0.17	0.59	0.12	233.4	84.3
24	My teacher explains things just as long until I get it	2.04	0.08	0.79	0.16	408.2	143.0
25	If my answer to a question is incorrect, my teacher explains why it is incorrect	2.77	0.10	0.76	0.16	388.4	127.5
Scale 4 – Challenging students							
26	My teacher wants me to do my best	−1.18	0.08	1.16	0.27	194.3	77.6
27	My teacher thinks I can do good work if I try my hardest	1.11	0.04	1.48	0.32	421.9	134.6
28	My teacher says we need to think carefully about how to do the assignments well	0.84	0.13	0.60	0.13	179.4	69.8
29	My teacher thinks I can learn everything if I do my best	1.76	0.06	1.30	0.24	456.5	183.0
30	My teacher is only satisfied when we do the best we can	−1.34	0.75	0.26	0.07	59.4	29.1
Scale 5 – Goal orientation							
31	My teacher tells us at the start of the lesson *what* we are learning	−3.53	0.15	0.64	0.14	237.8	108.2
32	My teacher tells us at the start of the lesson *why* we are learning	0.97	0.07	0.89	0.18	551.1	227.6
33	When my teacher marks my work, he/she writes on my papers to help me understand	1.32	0.12	0.65	0.13	357.0	151.1
34	My teacher asks at the end of the lesson what we have learned	1.25	0.03	1.60	0.33	1271.0	518.3
35	My teacher reminds us at the beginning of the lesson what we covered in the previous lesson	1.80	0.07	0.94	0.19	621.5	247.9
36	My teacher wants me to think carefully whether my answer is correct	−3.13	0.08	0.65	0.14	239.2	106.9

Note: d represents the location parameter, a the discrimination parameter.

Items with a high Information value contribute much to the reliability of that scale, which means that the value added of the items for the scale is high. Items with a low Information value do not contribute much to the reliability of the scale. As shown in Table 6, each item contributes to its scale; however, some more than others. This means that all items are related to their scale to some extent, but that some items are either very simple or do not discriminate between teachers, or both.

For each scale, items can be identified that contribute much to the reliability of the scale. This means that for Scale 1 (classroom climate) Item 1 ("My teacher is nice to me when I ask questions") is most informative, whereas for Scale 2 (classroom management) Item 9 ("When we are working individually, it is quiet in the classroom") is most informative. "My teacher knows when I understand something, and when I do not" is most indicative for the instruction scale (Scale 3). For the challenging students scale (Scale 4), the item "My teacher thinks I can learn everything if I do my best" is most informative, whereas Item 34 ("My teacher asks at the end of the lesson what we have learned") is most informative for the goal-orientation scale (Scale 5).

Comparing observers and students

To answer the third research question, the correlations between the estimates of teachers' teaching quality from the observer scales and the estimates from the student perception scales are presented in Table 7 (see Formula [1]). The approach used here is somewhat related to the multitrait-multimethod approach by Campbell and Fiske (1959) for assessing convergent and discriminant validity by inspecting patterns of correlations within and between scales. The relation between constructs and measures is less clear-cut in the present case, but the idea of looking for patterns remains the same.

Table 7. Correlation matrix for the observer scales (O1–O6) and the student perception scales (S1–S5).

	S1 – Classroom climate	S2 – Classroom Management	S3 – Instruction	S4 – Challenging students	S5 – Goal orientation
O1 – Classroom climate	.45	.50	.46	.32	.23
O2 – Classroom management	.35	.42	.43	.26	.21
O3 – Instruction	.43	.47	.48	.32	.26
O4 – Adapting instruction	.35	.30	.47	.26	.39
O5 – Learning strategies	.42	.42	.47	.30	.26
O6 – Engagement students	.35	.42	.41	.26	.18

As is presented in Table 7, the correlations between the student scales and the observer scales range from very low to moderate (r = .18 to r = .50). The three scales with comparable contents (Scales 1 to 3) all show low correlations (ranging from r = .42 to r = .48) but are higher than most of the other correlations. These subscales all measure teachers' behavior, which can be operationalized quite well for both external observers and students. However, the highest correlation was found for two unrelated scales: the classroom students' classroom management scale and the

observers' classroom climate scale (r = .50). A clear pattern could therefore not be identified. Nevertheless, correlations between, on the one hand, the student scales for instruction and, on the other hand, the instruction, adaption of instruction, and the learning strategies scales for the external observers are relatively high (r = .47 to r = .48). These lesson aspects mainly take place during the instruction phase of the lesson.

The student goal-orientation scale (S5) hardly correlates with any of the observer scales. Only the correlation with the adaption of instruction scale used by the observers is reasonable (r = .39), wherein two items from the adaption of instruction scale relate to discussing lesson goals.

The correlations that were found between the "Engagement of student" scale (O6) and the student perception scales were moderate to low (ranging from r = .18 to r = .42). A moderate correlation was found for the classroom management scale (r = .41) and the instruction scale (r = .42), indicating that when students are more engaged according to the observers, students report better instruction and classroom management.

Conclusion and discussion

In this study, we investigated the validity and reliability of student perceptions of teaching quality in primary education and their consistency with the ratings from by external observers. We will first elaborate on the findings and strengths of this study. Thereafter, the limitations of this study and suggestions for further research are presented.

The first research question was answered by studying two related types of validity: construct validity and discriminant validity. The confirmatory factor analysis, which can be used as an indicator of construct validity (Shadish et al., 2002), showed that the assumed scales for both the observer ratings and the student perceptions were represented in the data. In other words, students are able to discriminate between different teaching quality constructs. If not, the unidimensional or dimensional model (in which the observer and student perception scales were not taken into account) would have resulted in a better fit. Furthermore, the correlations between the various student perceptions scales also indicated that Grade 4 students were able to distinguish between different teaching quality constructs. It is, of course, difficult to define exactly how low the latent correlations should be to assume discriminant validity. However, they are so low that the hypothesis that the five latent dimensions form a unidimensional scale can clearly be rejected. In this respect, Brown (2015) gives a boundary correlation of .85. Since the correlations range between .42 and .74, unidimensionality is clearly rejected and the latent dimensions are distinguishable. This finding is very valuable for the use of student perceptions in primary education and is in line with the findings from other recent studies, for example, the studies by Kane, McCaffrey, and Staiger (2010), Kunter and Baumert (2006), and Fauth et al. (2014).

As far as the second research question is concerned, the global (scale-level) reliability and the local (item-level) reliability of the student perceptions measures were investigated. The

scales showed a high global reliability (even at the same level as the reliabilities of the observer scales), and the local reliability showed that individual items belonged to their scale. Furthermore, when considering the content of the most informative items per scale, each of these items was clearly related to the scale's content. For example, in the theoretical framework of this study we described that classroom management is about having clear class rules and routines, preventing disruptive behavior, and having a well-organized and clear lesson structure (Day et al., 2008). It makes sense that, from a student perspective, the statement "Everybody pays attention when my teacher explains something" reflects the extent to which teachers have good classroom management skills. Also for the instruction scale, which for lesson observers includes items like "The teacher checks during the instruction whether students understood the subject matter or not", the most indicative student perception item was "My teacher knows when I understand something, and when I do not". These findings suggest that student perceptions may be used for evaluating teachers' distinct teaching qualities.

When the student perceptions results were compared with the ratings by external observers (Research question 3), we found that the teaching quality scales of both measurements correlated to a certain extent with each other, ranging from $r = .18$ to $r = .50$, with 23 out of 30 correlations at or above $r = .30$. Considering the correlations found in previous studies, as described in the theoretical framework by De Jong and Westerhof (2001) and Maulana and Helmz-Lorenz (2016), it can be concluded that relatively high correlations were found in this study. However, the results in the other studies are not entirely comparable with the results found in the current study for several reasons. First, the studies were executed in different contexts (in secondary schools instead of in primary schools). Second, in the study of De Jong and Westerhof (2001), three lessons per teacher were observed (similar to the current study), but only one observer rated the lessons. Maulana and Helms-Lorenz (2016) observed only one lesson per teacher (rated by one observer). The current study incorporated three lessons and four raters, which is important for obtaining a reliable estimate of a teacher's teaching quality (Hill et al., 2012; Murray, 1983). Third, when student and observers' ratings were compared by De Jong and Westerhof (2001), they compared completely different teaching constructs, whereas in the current study it was intended to measure the three basic constructs of teaching quality in a similar way using both the student perception questionnaire and the observation scheme.

Despite the methodological strengths of this study, the correlations between students and observers still range from moderate to low. Although the correlations between the three scales with a comparable content (Scales 1 to 3) are higher than most of the other correlations, the highest correlation was found for two constructs that were not comparable. Therefore, a clear pattern could not be identified. This could be explained by the differences at the item level between both instruments (Maulana & Helms-Lorenz, 2016), where it was intended to measure the same construct. The items were not phrased identically across the observer and the student perceptions instruments, and the number of items per construct also differed across the measures.

Another explanation for the moderate to low correlations could be that, due to the differences in background and knowledge between external observers and students (as

mentioned in our theoretical framework), students and observers perceive the same constructs in a different way. This was, for example, found for the student perception instruction scale, on the one hand, and for the observers teaching learning strategies scale (r = .48) and the observers adaption of instruction scale (r = .47), on the other hand. This indicates that external observers and students observe similar aspects of teaching during the lesson. Again, this assumes that student perceptions might have the potential to be used for evaluating teaching quality (for professional development or research purposes) instead of using expensive and time-consuming classroom observations. However, as the relation between student perceptions and observations did not show a clear pattern, further in-depth research is required to gain more insight into this relationship.

Limitations of the study and recommendations for further research

A first remark related to the limitations of the study is that, although the same teacher was assessed by both observers and students, both measurements were conducted at different times during the school year. The majority of the lesson recordings were made in the year preceding the school year in which the student perception questionnaires were assessed. This may have affected the comparability of both measurements, and could have impacted the extent to which student perceptions and lesson observations were comparable.

Furthermore, in this study students provided their opinion about their teacher once, while three lessons were used to estimate teaching quality by means of observations. For future research, it is suggested to investigate the comparability of both measurements in a more strict way: by comparing teaching quality based on student perceptions and ratings by external observers for the *same lessons*, using the same wording and rating scales for the items.

Another suggestion for future research is to gain more insight into how students interpret the statements in the items of a student perception questionnaire, such that differences between what raters observe and students perceive can be indicated. This information could be used to adapt the questionnaires, such that the content is more aligned to the observation scheme. In addition, some items of the student perception questionnaire in this study did not provide much information and might be excluded in future research.

A final remark pertains to the statistical methodology used for this study for both student perception research and lesson observation research in general. The main objective of statistical modeling is to identify the different sources of uncertainty in order to develop reliable unbiased estimates of teaching proficiency. It is now generally accepted that the nested structure of the data should be addressed using multilevel models. Furthermore, in analyses of educational surveys, the use of an IRT model to model item responses is generally accepted. The reason is that just summing up item responses ignores the unreliability of the responses and does not differentiate between the items by weighting the responses, as is done in IRT. In this article, this approach is generalized by applying IRT modeling to the item responses of the student perception questionnaire and the observation instrument. We argue that this approach should be the standard in student perception and lesson observations research.

In summary, the results of this study are valuable for how we can use student perceptions of teaching quality for evaluating teaching quality and for feeding back the results of the evaluations to teachers. In response to this feedback, teachers can reflect on their lessons, and they may use the feedback to improve their teaching quality. Our research also points to new directions for future research which can help us gain an in-depth and precise understanding of what we measure when we study student perceptions of teaching quality and how we can utilize how students perceive the teaching quality of their teachers.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Emmelien A. van der Scheer http://orcid.org/0000-0003-2800-4698
Hannah J. E. Bijlsma http://orcid.org/0000-0001-8825-0167

References

Arends, R. I. (2009). *Learning to teach* (8th ed.). New York, NY: McGraw-Hill.
Ben-Chaim, D., & Zoller, U. (2001). Self-perception versus students' perception of teachers' personal style in college science and mathematics courses. *Research in Science Education*, *31*(3), 437–454. doi:10.1023/A:1013172329170
Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice*, *11*(4), 27–34. doi:10.1111/j.1745-3992.1992.tb00260.x
Brennan, R. L., & Johnson, E. G. (1995). Generalizability of performance assessments. *Educational Measurement: Issues and Practice*, *14*(4), 9–12. doi:10.1111/j.1745-3992.1995.tb00882.x
Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York, NY: The Guilford Press.
Burniske, J., & Meibaum, D. L. (2012). *The use of student perceptual data as a measure of teaching effectiveness*. Retrieved from the Texas Comprehensive Center website: http://txcc.sedl.org/resources/briefs/number_8/index.php
Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. doi:10.1037/h0046016
Creemers, B. P. M. (1994). *The effective classroom*. London: Cassell.
Day, C., Sammons, P., Kington, A., Regan, E., Ko, J., Brown, E., … Robertson, D. (2008). *Effective classroom practice (ECP): A mixed-method study of influences and outcomes*. Nottingham: School of education, University of Nottingham.
De Jong, R., & Westerhof, K. J. (2001). The quality of student ratings of teacher behaviour. *Learning Environments Research*, *4*(1), 51–85. doi:10.1023/A:1011402608575
Den Brok, P., Brekelmans, M., Levy, J., & Wubbels, T. (2002). Diagnosing and improving the quality of teachers' interpersonal behavior. *International Journal of Educational Management*, *16*(4), 176–184. doi:10.1108/09513540210432155
Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction*, *29*, 1–9. doi:10.1016/j.learninstruc.2013.07.001
Ferguson, R. F. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan*, *94*(3), 24–28. doi:10.1177/003172171209400306

Ferguson, R. F., & Danielson, C. (2014). How framework for teaching and Tripod 7Cs evidence distinguish key components of effective teaching. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems* (pp. 98–143). San Francisco, CA: Jossey-Bass.

Fraser, B. J. (1998). Science learning environments: Assessment, effects and determinants. In B. J. Fraser & K. G. Tobin (Eds.), *International handbook of science education* (pp. 527–564). London: Kluwer Academic.

Gaertner, H. (2014). Effects of student feedback as a method of self-evaluating the quality of teaching. *Studies in Educational Evaluation*, *42*, 91–99. doi:10.1016/j.stueduc.2014.04.003

Glas, C. A. W. (2012). Generalizability theory and item response theory. In T. J. H. M. Eggen & B. P. Veldkamp (Eds.), *Psychometrics in practice at RCEC* (pp. 1–13). Retrieved from https://ris.utwente.nl/ws/portalfiles/portal/5573389/Chapter_1.pdf

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist*, *52*(11), 1182–1186. doi:10.1037/0003-066X.52.11.1182

Hamre, B. K., & Pianta, R. C. (2010). Classroom environments and developmental processes: Conceptualization and measurement. In J. L. Meece & J. S. Eccles (Eds.), *Handbook of research on schools, schooling and human development* (pp. 25–41). New York, NY: Routledge.

Hattie, J. (2008). *Visible learning: A synthesis over 800 meta-analyses relating to achievement*. Abingdon: Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. doi: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. doi:10.3102/0013189X12437203

Hollingsworth, J., & Ybarra, S. (2009). *Explicit direct instruction (EDI): The power of the well-crafted, well-taught lesson*. Fowler, CA: DataWORKS Educational Research.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers?* Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/MET_Validating_Using_Random_Assignment_Research_Paper.pdf

Kane, T. J., McCaffrey, D. F., & Staiger, D. O. (2010). *Learning about teaching: Initial findings from the Measures of Effective Teaching project*. Seattle, WA: Bill & Melinda Gates Foundation.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains* (MET project Research Paper). Seattle, WA: Bill & Melinda Gates Foundation.

Keuning, T., & Van Geel, M. (2016). *Implementation and effects of a schoolwide data-based decision making intervention: A large-scale study*. Enschede: University of Twente.

Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environment Research*, *9*(3), 231–251. doi:10.1007/s10984-006-9015-7

Kyriakides, L. (2005). Drawing from teacher effectiveness research and research into teacher interpersonal behaviour to establish a teacher evaluation system: A study on the use of student ratings to evaluate teacher behaviour. *Journal of Classroom Interaction*, *40*(2), 44–66. Retrieved from http://www.jstor.org/stable/23870663

Kyriakides, L., Campbell, R. J., & Gagatsis, A. (2000). The significance of the classroom effect in primary schools: An application of Creemers' comprehensive model of educational effectiveness. *School Effectiveness and School Improvement*, *11*(4), 501–529. doi:10.1076/sesi.11.4.501.3560

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, *57*(9), 705–717. doi:10.1037/0003-066X.57.9.705

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

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

Maulana, R., Helms-Lorenz, M., & Van de Grift, W. (2014). Development and evaluation of a questionnaire measuring pre-service teachers' teaching behaviour: A Rasch modelling approach. *School Effectiveness and School Improvement*, *26*(2), 169–194. doi:10.1080/09243453.2014.939198

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

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176. doi:10.1177/014662169201600206

Murray, H. G. (1983). Low-inference classroom teaching behaviors and student ratings of college teaching effectiveness. *Journal of Educational Psychology*, *75*(1), 138–149. doi:10.1037/0022-0663.75.1.138

Patton, M. Q. (1980). *Qualitative evaluation methods*. Beverly Hills, CA: Sage.

Peterson, K. D., Wahlquist, C., & Bone, K. (2000). Student surveys for school teacher evaluation. *Journal of Personnel Evaluation in Education*, *14*(2), 135–153. doi:10.1023/A:1008102519702

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

Reynolds, D., Sammons, P., De Fraine, B., Van Damme, J., Towsend, T., Teddlie, C., & Stringfield, S. (2014). Educational effectiveness research (EER): A state-of-the-art review. *School Effectiveness and School Improvement*, *25*(2), 197–230. doi:10.1080/09243453.2014.885450

Rosenshine, B. (1995). Advances in research on instruction. *The Journal of Educational Research*, *88*(5), 262–268. doi:10.1080/00220671.1995.9941309

Rutkowski, L., von Davier, M., & Rutkowski, D. (Eds.). (2014). *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*. Boca Raton, FL: CRC Press.

Sammons, P., Hillman, J., & Mortimore, P. (1995). *Key characteristics of effective schools: A review of school effectiveness research*. London: Institute of Education.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Belmont, CA: Wadsworth.

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393–408. doi:10.1007/BF02294363

Timperley, H., Wilson, A., Barrar, H., & Fung, I. (2007). *Teacher professional learning and development: Best evidence synthesis iteration*. Wellington, New Zealand: Ministery of Education. Retrieved from http://www.educationcounts.govt.nz/__data/assets/pdf_file/0017/16901/TPLandDBESentireWeb.pdf

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

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

Van de Grift, W., Van der Wal, M., & Torenbeek, M. (2011). Ontwikkeling in de pedagogisch didactische vaardigheid van leraren in het basisonderwijs [Development of teachers' teaching skills in primary education]. *Pedagogische Studiën*, *88*(6), 416–432.

Van der Lans, R. M. (2017). *Teacher evaluation through observation: Application of classroom observation and student ratings to improve teaching effectiveness in classrooms*. Groningen: University of Groningen.

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. doi:10.1111/emip.12078

Van der Scheer, E. (2016). *Data-based decision making put to the test*. Enschede: University of Twente.

Van der Scheer, E. A., Glas, C. A. W., & Visscher, A. J. (2017). Changes in teachers' instructional skills during an intensive data-based decision making intervention. *Teaching and Teacher Education*, *65*, 171–182. doi:10.1016/j.tate.2017.02.018

Van der Scheer, E. A., & Visscher, A. J. (2018). Effects of a data-based decision-making intervention for teachers on students' mathematical achievement. *Journal of Teacher Education*, *69*(3), 307–320. doi:10.1177/0022487117704170

Wagner, W., Göllner, R., Helmke, A., Trautwein, U., & Lüdtke, O. (2013). Construct validity of student perceptions of instructional quality is high, but not perfect: Dimensionality and generalizability of domain-independent assessments. *Learning and Instruction*, *28*, 1–11. doi:10.1016/j.learninstruc.2013.03.003