

Chapter 2

A Reflection on Student Perceptions of Teaching Quality from Three Psychometric Perspectives: CCT, IRT and GT



Hannah Bijlsma, Rikkert van der Lans, Tim Mainhard, and Perry den Brok

Abstract This chapter discusses student perceptions in terms of three psychometric perspectives that dominate contemporary research on teaching quality, namely, Classical Test Theory (CTT), Item Response Theory (IRT) and Generalizability Theory (GT). These perspectives function as being exemplars for the connection between psychometric theories and the different perspectives on “what a perception is” as well as on how and for what purposes student perceptions should be used. The main message of the chapter is that the choice of a psychometric theory is not merely a technical matter, but also has implications for how the nature of perceptions is conceptualized. After presenting and linking each psychometric theory, their strengths and weaknesses in the context of student perceptions of teaching quality and issues on practical implementations are discussed.

Keywords Student perceptions · Teaching quality · Classical test theory · Item response theory · Generalizability theory

H. Bijlsma (✉)

Section of Teacher Professionalization, University of Twente,
Enschede, the Netherlands

e-mail: h.j.e.bijlsma@utwente.nl

R. van der Lans

Curium LUMC, Leiden, The Netherlands

e-mail: r.m.van_der_lans@curium.nl

T. Mainhard

Department of Education, University of Utrecht, Utrecht, The Netherlands

e-mail: m.t.mainhard@uu.nl

P. den Brok

Department of Educational and Learning Sciences, University of
Wageningen, Wageningen, The Netherlands

e-mail: perry.denbrok@wur.nl

© The Author(s) 2021

W. Rollett et al. (eds.), *Student Feedback on Teaching in Schools*,
https://doi.org/10.1007/978-3-030-75150-0_2

1 Introduction

Student perceptions of teachers and their behaviours have become an important way to capture what happens in class. Questionnaires that map student perceptions of teaching quality are used, for example, to measure the effectiveness of educational interventions (Burniske & Meibaum, 2012; Kyriakides, 2005). In schools, student perceptions are collected by teachers to obtain feedback for improvement and professional development activities (Bijlsma et al., 2019).

Using student perceptions of teaching quality is a complex process. Typically, perceptions are collected using a standardized questionnaire instrument. When a student selects a response category of an item like “my teacher explains everything clearly to me”, however, many processes may affect the student’s answer. For example, a student may deliberately give a higher rating for the item than their real estimation of their teacher’s skill at explanation because (s)he wants to present him/herself in a socially desirable way, or the student’s perception may be biased by stereotypical impressions. Alternatively, the student might be honest and their perception unbiased, but a misinterpretation of the item content, for example, a different interpretation of what clarity means in this context, may still affect the item response (Maulana & Helms-Lorenz, 2016).

Moreover, items can be formulated according to the level of behaviour at which they are directed (to an individual student or the whole class), and in terms of the level of perception (personal, class). In Chap. 7 by Göllner et al. in this volume, it is referred to as differences in the referent and in the addressee of items. For example, the aforementioned item can be worded as: “This teacher explains things clearly to us/the class” (class perception, behaviour to class), “This teacher explains things clearly to me” (class perception, behaviour to individual), “I find this teacher to explain things clearly” (personal perception, behaviour to class) and “I find this teacher to explain things clearly to me” (personal perception, behaviour to individual). While this may seem trivial, it has consequences for the expected sources of variation in perceptions: items asking about class perceptions or behaviours directed at the whole class are more likely to evoke variation in shared sources of perceptions, while items asking about behaviours directed at individuals or personal perceptions are more likely to evoke variation in idiosyncratic sources of perceptions.

The question of what we actually measure, therefore, has no uniform answer. By completing standardized questionnaires, students give responses to many items and psychometric models are applied to combine the item ratings into an overall student perception of teachers’ teaching (students’ responses are then combined to a numerical value or score). This overall score—not the item ratings—is usually fed back to teachers or is used for research purposes. This approach of combining and integrating ratings into one overall perception score suggests that students cognitively process observations of teaching behaviours similarly and in such a general and integrated way. From this perspective, the psychometric models that connect and integrate the item ratings attempt to reconstruct students’ mental representations of the teachers’ teaching.

This chapter discusses student perceptions in terms of three psychometric perspectives that dominate contemporary research on teaching quality, namely, Classical Test Theory (CTT), Item Response Theory (IRT) and Generalizability Theory (GT). CTT (part 2) is based on the assumption that there is one true score and a variance score (error). The true score is then an average of all students' ratings on certain items that form a dimension or factor. In IRT (part 3), more emphasis is put on how many items relate to each other and what dimensions can be distinguished in the instrument used to collect student perceptions of teaching quality. The potential of GT (part 4) lies in the fact that it tries to disentangle the variability in student ratings beyond a "true score" and error, bringing in aspects such as personal characteristics and dyadic relationships between people. The chapter discusses these psychometric perspectives separately, but there are also integrated approaches that can enable researchers to estimate combinations of the models (Chalmers, 2012; Robitzsch et al., 2020). The connection between the CTT, IRT and GT with latent variable models becomes evident when it is realized that all specify a relationship between the teachers' latent ability level and the responses of students that were stimulated (or elicited) by the items (e.g., Chalmers, 2012; de Boeck et al., 2011; Rizopoulos, 2006; Robitzsch et al., 2020).

The main message of the chapter is that the choice of a psychometric model is not merely a technical matter, but also has implications for how the nature of perceptions is conceptualized. Finally, we acknowledge that the construct of teaching quality is highly contested and consensus about its conceptualization or definition is minimal (Cohen & Goldhaber, 2016). We do not present a definition of teaching quality in this chapter. By leaving the definition completely open, we intend to maximize our flexibility to discuss various possibilities offered by the three psychometric theories. After presenting and linking each psychometric theory, we will discuss their strengths and weaknesses in the context of student perceptions of teaching quality.

2 Classical Test Theory

2.1 *The CTT Model*

According to Classical Test Theory (CTT), student perceptions of teaching quality reflect the teachers' actual teaching quality plus random error variance (e.g., Brennan, 2001; Lord & Novick, 1968; Sijtsma, 2016; Spearman, 1905). The teachers' actual teaching quality is caught by the so-called "true score", which is statistically defined by the mean score over all item responses about that teacher. The error variance consists of all random deviations from the teacher's mean score (Novick, 1966). Furthermore, the CTT model states that all items are equally associated with the broader perceptual representation of the teachers' teaching (i.e., items are supposed to have similar factor loadings).

Table 1 Possible example of feedback form results for one teacher teaching a class of 25 students

Item My Teacher...	N_{class}	Class mean	Class SD
...			
... makes sure that others treat me with respect.	25	3.28	0.52
... makes clear what I need to learn for a test.	25	3.14	0.78
... explains everything clearly to me.	25	2.72	0.94
... uses clear examples.	25	2.82	0.93
... encourages me to cooperate with my classmates.	25	2.08	0.80
...			
Total	25	2.81	0.79

Marsh (2007) noted that overall questionnaire outcomes may be uninformative about *specific* teaching behaviours, and therefore recommends structuring questionnaires according to different factors. Factors cluster items that seem to have something in common based on the inter-item correlations. For example, the items, “My teacher explains everything clearly to me” and “My teacher uses clear examples” (see Table 1), are connected to the same factor, which clusters items related to the clarity and structuredness of explanations (Maulana & Helms-Lorenz, 2016). Reporting the class mean for items related to the clarity and structuredness of explanations is considered more informative than just an overall mean for all items.

In educational contexts, the CTT model is usually extended by including multiple nested levels of random error; for example, students are nested within teachers. The key idea of CTT, however, remains, in that only the mean of a factor is informative and variation around the mean is uninformative noise.

Paramount to the logic behind CTT is that item ratings related to the same teacher should show minimal variability *and* that item ratings related to different teachers should show large(r) variation. Hence, item ratings assigned by one student to the same teacher are expected to vary minimally. The mean student questionnaire scores from students within the same class are also expected to show minimal variability. These expectations are routinely examined by estimates of internal consistency (Cronbach, 1951) and intra-class correlations (ICCs; Lüdtke et al., 2009). Internal consistency is sensitive for items showing large variation in ratings compared to the other items’ ratings. The ICC provides an estimate of the variance in mean questionnaire scores from students in different classes as proportionate to the variance of all ratings.

2.2 *An Example of CTT in Practice*

Suppose that 25 students in a class respond to the item “My teacher explains everything clearly to me” by choosing one of the four answer options: 1 = “never”, 2 = “seldom”, 3 = “occasionally” and 4 = “often” (Table 1). If CTT is applied strictly, then the mean class perception (2.72) is the only reliable and, thus, the only informative parameter for the teacher to consider, and individual deviations are *random* noise. This logic can easily be generalized to a broader set of items. For example, the mean of the student questionnaire ratings can be computed and CTT can be applied to these mean scores, which may then be argued to be the most reliable estimate of the teacher’s actual teaching quality. In this example, according to CCT, 2.81 reflects the teacher’s teaching quality based on these five items.

2.3 *Advantages and Limitations of the CTT Approach*

The CTT approach, and Marsh’s (1987, 2007) work in particular, are well-known and studied in the educational sciences. Estimates of internal consistency and ICCs have proven to be stable across different questionnaires (cf. Marsh, 2007; van der Lans & Maulana, 2018). These statistics are also intuitively understandable for many practitioners and the application of CTT requires only a modest level of mathematical and statistical skill, which is not unimportant.

However, the use of CTT reflects high trust in the students as being honest and accurate perceivers. To illustrate this, suppose that students deliberately manipulate their ratings upwards because they like the teacher; then clearly such *systematic* bias or manipulation remains undetected by measures such as internal consistency and ICC, which quantify *random* error variance only (den Brok & Smart, 2007). In general, CTT provides very limited means to empirically investigate systematic biases in perceptions. Second, diagnosing poor item quality by the comparatively large variance in ratings, as is done by internal consistency measures, is only valid if one believes that ratings of all items must be biased by the same amount of (random) error. Suppose again that students deliberately manipulate their ratings upwards because they like the teacher; then their manipulation might well be expressed most in items referring to specific teacher traits that are likable (such as “humour”, or “showing respect”). More in general, CTT fails to make (differentiated) predictions about the response process; for example, when students check a response category, it remains unsolved what latent cognitive representation of the teacher’s teaching students had in mind.

3 Item Response Theory

3.1 Item Response Theory (IRT) Model(s)

According to IRT, student perceptions of teaching are ordered on a latent continuum (Bond & Fox, 2007; Embretson & Reise, 2013). With IRT, researchers estimate the teacher’s position on this latent continuum and this position is then used to predict the most likely teacher behaviour that students will have experienced from this teacher. There are two levels at which IRT can be used to make predictions about what teacher behaviours students likely will have experienced: (1) the level of the item and (2) the level of the construct (Bond & Fox, 2007; Embretson & Reise, 2013). At the level of the item, IRT uses the response categories to make predictions about whether students experienced that particular behaviour seldom, occasionally or often. At the level of the construct, IRT makes predictions about how items jointly represent the teachers’ teaching.

We will explain this by using one of the five items from Table 1 (“My teacher explains everything clearly to me” [explains clearly]). In Fig. 1, the y-axis indicates the probability of checking the higher response category out of two competing response categories and the x-axis indicates the level of teaching quality (θ). Teachers with a level of teaching quality located at the position of the arrow have a high probability of receiving a response “ \geq seldom” on explains everything clearly to me, but a low probability of receiving a response “occasionally”. The probability that students check the higher response category increases only when the teacher—according to the responding student—has achieved the conditions set by the higher response category for the item.

The item response process can be used to predict the most likely frequency with which the behaviour is observed (or the most likely impact, if the item labels are insufficient, sufficient, excellent). This item response process is part of a wider process here referred to as the construct response process. The construct response process predicts how students weigh and position items relative to other items. In IRT, one

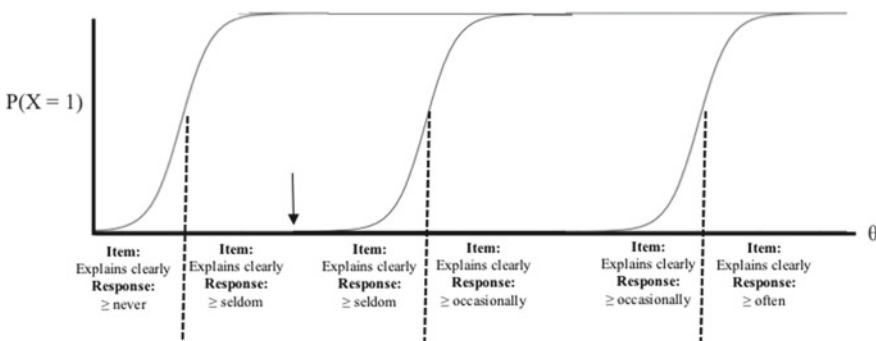


Fig. 1 Visualization of the item and construct response process

Fig. 2 The Guttman scale/simplex construct response (pattern obtained from: Mokken et al., 2001)

Guttman scale or simplex					
Item	1	2	3	4	5
Student A	✓				
Student B	✓	✓			
Student C	✓	✓	✓		
Student D	✓	✓	✓	✓	
Student E	✓	✓	✓	✓	✓
Student F	✓	✓	✓	✓	✓

well-known construct response process is the Guttman scale or simplex¹ (Guttman, 1954; Jöreskog, 1978). In the simplex, item positions depend on their “difficulty”. Some items are much more likely to receive the rating “never” (called “difficult” items), while other items are much more likely to be rated as “often” (the “easy” items). Figure 2 visualizes this pattern using five items. In Fig. 2, the checkmarks indicate a high probability that students perceive the teacher to perform the behaviour described by the item often. Hence, student D is predicted to perceive the teacher as performing the first four behaviours often, but not the fifth. Item 1 would be a “difficult” item, and 5 would be an “easy” item.

To order items, IRT models include a location parameter (sometimes referred to as item difficulty). The location parameter predicts when the item response process changes within the wider construct response process. For example, the response process for item four is predicted to change if the first three items have received high ratings. Other item parameters that can be estimated by IRT models are the discrimination parameter (to predict and correct for systematic deviations from the predicted item response process), and a guessing parameter (to predict and correct for randomness in the item response process). In what follows, we will present an example of research applying IRT to student perceptions to illustrate the above.

3.2 IRT in Research on Student Perceptions

Van de Grift and Kyriakides started independently implementing IRT in the context of teaching quality with student perception data (for details, see Antoniou & Kyriakides, 2013; Kyriakides et al., 2018; Maulana et al., 2015; van de Grift et al., 2011, 2014; van der Lans et al., 2015). Their models hypothesize that teaching effectiveness develops along a latent continuum in which learning to teach starts with learning less complex teaching behaviours (e.g., ensuring a safe classroom climate) and ends

¹ There are two other main classes of construct response processes, namely the Coombs/unfolding and the circumplex (Browne, 1992; de Leeuw & Mair, 2011; Mokken et al., 2001). It goes beyond the scope of this chapter to define and describe these as well. Hence, we focus here on the Guttman scale/simplex construct response.

with learning more complex teaching behaviours (e.g., having students cooperate with classmates). Hence, students' ratings given on questionnaires that list various teaching behaviours should indicate that they perceive some more complex teaching behaviours to be performed successfully less frequently, whereas they perceive other less complex teaching behaviours to be performed successfully more often (by more teachers). These researchers have applied Rasch-family models²—a specific type of IRT model—to test sequences of item complexity and to locate teachers on the latent continuum. After they have located the teacher, they provide the teacher with feedback by indicating the next steps for improvement (i.e., the items located just beyond the teacher's position). In other recent research, IRT has been used to examine issues of validity of student perception data (e.g., Bijlsma et al., submitted; van der Scheer et al., 2018).

3.3 Advantages and Limitations of IRT Models

The comprehensive framework of IRT provides various possibilities for testing hypotheses concerning students' response processes at the level of the item and at the level of the construct. Thereby, IRT is promising as a way to develop and test theories that predict how different formulations of survey items and/or formulations of response categories translate into distinct item response and construct response processes. Substantive theories can also be translated into item and construct response processes, as in the example described in the previous section.

However, the disadvantage of IRT is that it basically assumes that the item response process is unbiased. Take the research we discussed by van de Grift et al. (2014). They predicted that student ratings will follow sequences predicted by theory on teacher development, but this prediction assumes that student ratings are a direct (unbiased) numerical representation of the teacher's actual behaviour. IRT can include a discrimination parameter to correct for systematic biases, but this discrimination parameter corrects the item response process for all biases and generally is uninformative about the potential sources of bias. Various biases will impact the students' item responses, such as social desirability and stereotypical views (Kenny, 1994). As we will detail next, generalizability theory provides a framework for examining such influences on item ratings.

² Rasch-family models are applied to test the theoretical models, because Rasch model fit tests were developed to empirically examine hierarchical orderings in item ratings (Bond & Fox, 2007). Hence, if student perceptions are unbiased, then their responses could be used to locate the teacher on this latent novice–expert continuum.

4 Generalizability Theory

Generalizability Theory (GT) extends Classical Test Theory (CTT) by introducing the possibility of including systematic variance components (or facets) other than error and a teacher's "true score" (Brennan, 2001). The basic idea is that what is called error in CTT can be further sub-divided into systematic facets or sources of variability (Malloy, 2018) that potentially affect student perceptions of teaching quality. When such variance components are considered nuisance parameters, GT conceptually coincides with CTT, as it is viewed by Marsh (2007), for example. Traditionally, in the educational context, GT has been used to determine the number of tasks or raters that yield reliable test results (Shavelson & Webb, 2005). As such, the amount of error that tasks introduce or the degree of consensus between raters is typically GT's main focus. Yet, the strength of GT is that it can also be used to embrace and study 'error' in an attempt to learn more about how these additional sources of variability impact perceptions of social phenomena such as teaching.

4.1 A Practical Example Using GT and Student Ratings

One of the best-known models in social science that applies GT to social perceptions and interactions is Kenny's Social Relations Model (SRM, 1994). The basic assumption of the SRM is that any rating of a social perception has, besides error, three potential sources: an actor or rater effect (i.e., due to the student who responds to an item), a partner or target effect (i.e., due to the teacher who is rated) and a relationship effect (variability introduced due to the specific combination of this student rating that specific teacher). The partner or target effect resembles what is taken to be the teacher's true score or true ability in CTT. The variance in partner effects captures the degree of consensus between students on a certain aspect of teaching quality. Stable response tendencies within students are captured in the actor effect. For example, some students are quick learners and may therefore readily indicate that they understand teacher explanations, irrespective of a specific teacher's quality. There can also be systematic variance in ratings due to the relationship between, or the specific pairing of, students and teachers. Thus, on top of a student's stable tendency to think that teachers can explain things well (rater effect) and the teacher's general ability to explain things (target effect), student A may have experienced instances where teacher B has explained content exceptionally well. This shared interaction history may affect student A's ratings over and above the rater and target effects (Mainhard et al., 2018).

GT and SRM can be applied at the item level, though they are more commonly applied at the construct level (Kenny, 1994, 1996; Kenny et al., 2006). Let us consider an example at the item level. Suppose that students complete the item "my teacher explains everything clearly to me"; then at the item level, the SRM is informative: about the target effect, namely, do students agree that some teachers explain things

well while others are not? about the actor effect, namely, do some students tend to experience all teachers' explanations to be clear while other students tend to perceive all teachers' explanations as hard to understand? and about the dyadic effects, namely, do some students experience the teacher's explanation to be clear over and above their personal actor effect and that teacher's target effect? Note that the actor and relationship variances would be considered as error in CTT. The variability found in these sources can then be explained with predictors, as in regression analysis. For example, students' actor effects may be explained by their general academic ability and teachers' target effects by years of experience. Relationship effects may occur, for example, because some teachers think that certain students require a certain kind of explanation to understand the subject matter.

4.2 Advantages and Limitations of Generalizability Theory

An advantage of dealing with student ratings of teaching quality according to the GT approach is that it is a relatively simple extension of the better-known CTT. Those acquainted with multilevel analyses will find GT quite straightforward (Kenny et al., 2006). Conceptually, GT is more informative about potential variables that impact students' item responses. When items barely show stable variance between students, the responses are only minimally affected by students' personal characteristics and answer tendencies.

However, compared to IRT, GT puts little emphasis on how item ratings can be organized into a broader representation of teaching. Like CTT, GT is applied to sets of items that have a similar association with the latent construct. Further, the GT approach requires complex data sets. It cannot be applied with datasets that pair one class with a teacher. Instead, students need to complete a questionnaire for several teachers, and teachers need to be rated by several classes (see Mainhard et al., 2018 for an example).

5 Discussion

In this chapter, three dominant psychometric theories were discussed within the domain of research on the validity and reliability of student perceptions of teaching quality: Classical Test Theory, Item Response Theory and Generalizability Theory. While each of these models has its specific advantages and disadvantages, together they shed more complete light on what constitutes and determines students' perceptions of teaching quality, disentangling true scores from error, and distinguishing between more systematic and more random sources of variation in perceptions. Together, they present a nuanced and complex picture of what makes a (student) perception, and also how it can be used in research.

The main message of the chapter is that the choice of a psychometric model is not merely a technical matter, but also has implications for how the nature of perceptions is conceptualized. For example, statistical techniques or software are tools that can be of help, but they depend on the specific theory about what teaching (quality) is and what dimensions or constructs and their interrelationships underlie such behaviour. Regardless of the three theories described in this chapter, many instruments measuring student perceptions are based on effectiveness research. It mainly includes variables that have been found to be associated with student outcomes in correlational research, rather than specifying a structure in and between different dimensions of teaching and their likelihood of (co-)occurring (Skourdoumbis & Gale, 2013; Wrigley, 2004). For this purpose, CTT can be applied. Furthermore, many instruments are based on the frequency of occurrence of behaviours, assuming an order or singular dimension in these occurrences that is based on difficulty, routine or other phenomena, which is linked to IRT (Maulana & Helms-Lorenz, 2016; den Brok et al., 2018). However, others have argued that teaching quality is multidimensional in nature, with behaviours being interpretable from various perspectives and adding value to different outcomes at the same time (Doyle, 1986; den Brok, 2001; den Brok et al., 2004; Shuell, 1996). GT can be applied here.

One may argue that basing a theory about teaching quality on the actual presence of behaviour or association with existing student outcomes is conservative, and does not allow exploration of new teaching methods, new organisational forms of education or alternative learning outcomes. However, assumptions behind the occurrence of behaviours may differ depending on the type of perspective taken on teaching, as may their theoretical underpinnings. For example, many interactional theories assume two independent dimensions behind teaching, that order components of behaviour in circumplex structures with specific patterns and interrelations between behaviours (or items) (Fabrigar et al., 1997; Gurtman & Pincus, 2000; Wubbels et al., 2006). The more specified theories are, the easier they can be tested statistically, as many programmes assume or ask for specific relations to be tested when studying perceptions; consider, for example, structural equation modelling, confirmatory factor analyses, IRT analysis or latent variable analysis (den Brok et al., 2018).

6 Putting it all Together

With this chapter, we hope to have provided more insight into the interesting, yet complicated, world of student perceptions of teaching quality. In conclusion, we have a few take-away messages for researchers interested in using student perceptions of teaching.

First, as aforementioned, it is important to be specific about the underlying assumptions one has about the nature of the student perceptions one is interested in. These assumptions should be grounded in prior research conducted on perceptions of the particular teaching behaviours one is interested in. For example, are the

perceptions expected to vary considerably between teachers, classes or schools? Are the perceptions likely to evoke certain psychological processes, such as social desirability or stereotypical responses? Are the behaviours expected to be familiar or unfamiliar to perceivers? Depending on what is known or deemed relevant, researchers can choose between one or several of the theories mentioned in this chapter.

Second, it is important to be specific about the wording of the items capturing the perceptions, as wording may lead to differences in response patterns, and thereby differences in sources of variance that may occur, related to either perceiver, object or the relation between them. Typically, researchers are not that conscious about the choices and assumptions they make about perceptions and the wording they use.

Third, it is important to conceptualize and make explicit the different dimensions or constructs one is interested in and the expected relationships between them, preferably based on theory (and empirical results). As this chapter has shown, constructs may relate to each other in terms of difficulty or chance of occurrence (as with simplex structures), but also in terms of relatedness or independence (as with circumplex structures).

When researchers take all of these reflections into account, interesting insights may be obtained by collecting student perceptions of teaching, and by comparing these with, for example, the perceptions of others, such as teachers themselves. The present chapter provides an overview of techniques and three major theories that may be used to analyse and conceptualize such perceptions.

References

- Antoniou, P., & Kyriakides, L. (2013). A dynamic integrated approach to teacher professional development: Impact and sustainability of the effects on improving teacher behaviour and student outcomes. *Teaching and Teacher Education*, 29(1), 1–12. <https://doi.org/10.1016/j.tate.2012.08.001>.
- Bijlsma, H. J. E., Glas, C. A. W., & Visscher, A. J. (submitted). The reliability and construct validity of student perceptions of teaching quality. *Educational Assessment, Evaluation and Accountability*.
- Bijlsma, H. J. E., Visscher, A. J., Dobbelaer, M. J., & Veldkamp, B. P. (2019). Does smartphone-assisted student feedback affect teaching quality? *Technology, Pedagogy & Education*. <https://doi.org/10.1080/1475939X.2019.1572534>.
- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Lawrence Erlbaum Associates.
- Brennan, R. L. (2001). *Generalizability theory: Statistics for social science and public policy*. Springer-Verlag. <https://doi.org/10.1007/978-1-4757-3456-0>.
- Browne, M. W. (1992). Circumplex models for correlation matrices. *Psychometrika*, 57(4), 469–497.
- Burniske, J., & Meibaum, D. (2012). *The use of student perception data as a measure of teaching effectiveness*. Texas Comprehensive Center.
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6). <https://doi.org/10.18637/jss.v048.i06>.

- Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher*, 45(6), 378–387. <https://doi.org/10.3102/0013189X16659442>.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334. <https://doi.org/10.1007/BF02310555>.
- de Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, 39(11), 1–28.
- de Leeuw, J., & Mair, P. (2011). *Multidimensional scaling using majorization*. SMACOF in R.
- den Brok, P. (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., & Smart, J. C. (Eds.). (2007). *The scholarship of teaching and learning in higher education: An evidence-based perspective*. Springer. <https://doi.org/10.1007/1-4020-5742-3>.
- den Brok, P., Brekelmans, M., & Wubbels, T. (2004). Interpersonal teacher behavior and student outcomes. *School Effectiveness & School Improvement*, 15, 407–442.
- den Brok, P., Wubbels, T., & Mainhard, T. (2018). Developments in quantitative methods and analyses to study learning environments. In D. Zandvliet & B. Fraser (Eds.), *Thirty years of learning environments: Looking back and looking forward: Advances in learning environments research, volume 11* (pp. 41–58). Brill/Sense.
- Doyle, W. (1986). Classroom organization and management. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 392–431). Macmillan.
- Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.
- Fabrigar, L. R., Visser, P. S., & Browne, M. W. (1997). Conceptual and methodological issues in testing the circumplex structure of data in personality and social psychology. *Personality and Social Psychology Review*, 1, 184–203.
- Gurtman, M. B., & Pincus, A. L. (2000). Interpersonal adjective scales: Confirmation of circumplex structure from multiple perspectives. *Personality and Social Psychology Bulletin*, 26, 374–384. <https://doi.org/10.1177/0146167200265009>.
- Guttman, L. L. (1954). A new approach to factor analysis: the radex. In P. F. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences*. The Free Press.
- Jöreskog, K. G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika*, 43(4), 443–477.
- Kenny, D. A. (1994). *Interpersonal perception: A social relation analysis*. Guilford.
- Kenny, D. A. (1996). Models of non-independence in Dyadic research. *Journal of Social and Personal Relationships*, 13(2), 279–294. <https://doi.org/10.1177/0265407596132007>.
- Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*. Guilford Press.
- 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.
- Kyriakides, L., Creemers, B. P., & Panayiotou, A. (2018). Using educational effectiveness research to promote quality of teaching: The contribution of the dynamic model. *ZDM Mathematics Education*, 50(3), 381–393.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley Publishing Company.
- Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology*, 34(2), 120–131. <https://doi.org/10.1016/j.cedpsych.2008.12.001>.
- Mainhard, M. T., Oudman, S., Hornstra, L., Bosker, R. J., & Goetz, T. (2018). Student emotions in class: The relative importance of teachers and their interpersonal relations with students. *Learning and Instruction*, 53, 109–119. <https://doi.org/10.1016/j.learninstruc.2017.07.011>.
- Malloy, T. E. (2018). *Social relations modeling of behavior in dyads and groups*. Academic Press.

- Marsh, H. D. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–383). Springer. https://doi.org/10.1007/1-4020-5742-3_9.
- Marsh, H. W. (1987). Students' evaluations of University teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research*, *11*(3), 253–388. [https://doi.org/10.1016/0883-0355\(87\)90001-2](https://doi.org/10.1016/0883-0355(87)90001-2).
- 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. <https://doi.org/10.1007/s10984-016-9215-8>.
- Maulana, R., Helms-Lorenz, M., & van de Grift, W. J. C. M. (2015). Development and evaluation of a questionnaire measuring pre-service teachers' teaching behaviour: A Rasch modelling approach. *School Effectiveness and School Improvement*, *26*(2), 169–194. <https://doi.org/10.1080/09243453.2014.939198>.
- Mokken, R. J., van Schuur, W. H., & Leeferink, A. J. (2001). The circles of our minds: A nonparametric IRT model for the circumplex. In *Essays on item response theory* (pp. 339–356). Springer. https://doi.org/10.1007/978-1-4613-0169-1_18.
- Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, *3*(1), 1–18.
- Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–25. <https://doi.org/10.18637/jss.v017.i05>.
- Robitzsch, A., Kiefer, T., Wu, M., Robitzsch, M. A., Adams, W., Rupp, L., et al. (2020). *Package 'TAM': Test analysis modules*. Version: 3: 4, 26.
- Shavelson, R. J., & Webb, N. M. (2005). *Generalizability theory*. <https://web.stanford.edu/dept/.../GTheoryAERA.pdf>.
- Shuell, T. J. (1996). Teaching and learning in a classroom context. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 726–764). Macmillan.
- Sijtsma, K. (2016). Classical test theory. In S. J. Henly (Ed.), *Routledge international handbook of advanced quantitative methods in nursing research* (pp. 29–43). Routledge.
- Skourdoumbis, A., & Gale, T. (2013). Classroom effectiveness research: A conceptual critique. *British Educational Research Journal*, *39*(5), 892–906. <https://doi.org/10.1002/berj.3008>.
- Spearman, C. (1905). "General intelligence" objectively determined and measured. *American Journal of Psychology*, *15*, 201–293.
- van de Grift, W. J. C. M., van der Wal, M., & Torenbeek, M. (2011). Ontwikkeling in de pedagogische didactische vaardigheid van leraren in het basisonderwijs [Development of pedagogical didactic skills of primary school teachers]. *Pedagogische Studiën*, *88*(6), 416–432.
- van de Grift, W., Helms-Lorenz, M., & Maulana, R. (2014). Teaching skills of student teachers: Calibration of an evaluation instrument and its value in predicting student academic engagement. *Studies in Educational Evaluation*, *43*, 150–159. <https://doi.org/10.1016/j.stueduc.2014.09.003>.
- van der Lans, R. M., & Maulana, R. (2018). The use of secondary school student ratings of their teacher's skillfulness for low-stake assessment and high-stake evaluation. *Studies in Educational Evaluation*, *58*, 112–121. <https://doi.org/10.1016/j.stueduc.2018.06.003>.
- van der Lans, R. M., van de Grift, W. J., & van Veen, K. (2015). Developing a teacher evaluation instrument to provide formative feedback using student ratings of teaching acts. *Educational Measurement: Issues and Practice*, *34*(3), 18–27. <https://doi.org/10.1111/emip.12078>.
- van der Scheer, E. A., Bijlsma, H. J. E., & Glas, C. (2018). Validity and reliability of student perceptions of teaching quality in primary education. *School Effectiveness and School Improvement*. <https://doi.org/10.1080/09243453.2018.1539015>.
- Wrigley, T. (2004). School effectiveness: The problem of reductionism. *British Educational Research Journal*, *30*(2), 227–244. <https://doi.org/10.1080/0141192042000195272>.
- Wubbels, T., Brekelmans, M., Den Brok, P., & Tartwijk, J. (2006). An interpersonal perspective on classroom management in secondary classrooms in the Netherlands. In C. Everson & C. Weinstein (Eds.), *Handbook of classroom management: Research, practice, and contemporary issues* (pp. 1161–1191). Lawrence Erlbaum Associates.

Hannah Bijlsma is a researcher (Ph.D.) at the section of Teacher Professionalization at the University of Twente (the Netherlands) and a primary school teacher (grade 1). Her research focuses on measuring teaching quality and on the use of student perceptions of teaching quality in school contexts. In 2016 she founded a professional association for academic primary school teachers, of which she has been chairman for about five years. She is now a board member of the International Congress for School Effectiveness and Improvement (ICSEI) and a board member of the EARLI SIG on School Effectiveness and Improvement.

Rikkert van der Lans is a postdoctoral researcher currently working at the Department of Child and Adolescent Psychiatry, Curium-Leiden (the Netherlands). Previously he worked as a post-doctoral researcher at the department of teacher education of the University of Groningen, as a lecturer in methods and statistics at the department of educational sciences (GION) of University of Groningen, and as lecturer in the field of psychometrics at the department of methods and statistics of the University of Tilburg (the Netherlands). His research focuses on the evaluation of professional development and the psychometric assessment of quality tests.

Tim Mainhard is Associate Professor at the Department of Education at Utrecht University (the Netherlands). His research focuses on social dynamics in educational settings—specifically teacher-student interactions and relationships—and their impact on student and teacher outcomes, such as emotions, motivation, and academic achievement. Tim teaches classroom management courses at the Utrecht Graduate School for Teaching in both the primary and secondary teacher education programs. Tim has been chair of the Classroom Management Special Interest Group of the American Educational Research Association and is an associate editor for the journal *Learning and Instruction*.

Perry den Brok is full Professor and chair of the Education and Learning Sciences group at Wageningen University and Research (the Netherlands), and chair of the 4TU Centre for Engineering Education. His research focuses on educational innovation in higher education, teacher learning and professional development, teacher-student interpersonal relationships, and educational learning environments—both in-class as well as out-of-school learning environments. He was European editor of the *Learning Environments Research* journal (Springer) for well over 10 years. He has published several review articles and book chapters on teacher effectiveness and teacher collaborative learning.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

