



Measuring university teachers' teaching quality: a Rasch modelling approach

Ine Noben¹ · Ridwan Maulana¹ · Jan F. Deinum¹ · W. H. Adriaan Hofman¹

Received: 29 November 2019 / Accepted: 30 May 2020 / Published online: 9 June 2020
© The Author(s) 2020

Abstract

Insights into how university teachers develop their teaching can strengthen the effectiveness of professional learning activities. Professional learning initiatives aim to support teachers in developing a teaching profile that is focused on student learning. However, university teachers often report a combination of content- and student-focused approaches to teaching and are not systematically focused on the learning outcomes of students. Teacher development theories hypothesise teaching growth as a hierarchy in which the development of more-complex skills (student-centred teaching) is based on the easier ones. This cumulative development of university teachers' teaching has not yet been examined in authentic learning environments. We explored whether the hierarchy can also be found in observed teaching behaviour. Rasch analysis of 203 classroom observations revealed a stage-wise ordering of the teaching skills from basic to complex, consistent with the theorised development of teaching in the literature. University teachers develop from teacher-centred to student-centred teaching by increasingly acknowledging the student in the teaching and learning process. At the same time, they aim first to gain comfort in a stage (self-focused) before improving their teaching in that development stage (focus on task). As teachers develop, they move on to student-learning focused skills, such as teaching–learning strategies.

Keywords Classroom observations · Rasch analysis · Teaching development · Teaching practices

Introduction

High-quality teaching in higher education is generally defined as teaching that fosters students' deep approach to learning. When students take up a deep approach to learning, they focus on analysing and understanding information by connecting ideas and using evidence (Postareff et al. 2018). To support university teachers in developing their teaching quality, higher-education institutions increasingly offer professional learning activities (Saroyan and Trigwell 2015). The underlying idea in encouraging academics to take part in

✉ Ine Noben
i.noben@rug.nl

¹ Faculty of Behavioural and Social Sciences, University of Groningen, Groningen, The Netherlands

professional learning initiatives is to support “a stage-based progression in [teaching] competence from novice to expert” (Webster-Wright 2009, p. 718). The initiatives are more likely to foster teaching growth when they are situated in a university teachers’ own teaching practice and when feedback is targeted to their ‘zone of proximal development’ (Nevgi and Löfström 2015; Saroyan and Trigwell 2015). In order to appropriately tailor feedback that supports growth (i.e. development) in teaching quality, it is important to understand how teachers develop their teaching (Van der Lans et al. 2017). Research to date has largely focused on studying teaching development through self-reports of teachers. However, university teachers’ teaching development has not yet been examined in terms of teachers’ teaching behaviour in authentic learning environments.

In the present study, we identified whether there is teaching behaviour development in the university setting by applying item response theory (i.e. Rasch modelling). We selected the Rasch model because it offers advantages over classical test theory for studying teaching behaviour. Cavanagh and Waugh (2011, p. xi) mentioned multiple advantages of Rasch modelling: “(1) producing linear, unidimensional scales; (2) requiring that data must fit the measurement model; (3) producing scale-free person measures; (4) producing sample-free item difficulties; (5) calculating standard errors; (6) estimating person measures and item difficulties on the same linear scale in standard units (logits); and (7) checking that the scoring system is being used logically and consistently”. These properties make it possible to identify the development of teaching behaviour in terms of complexity levels, taking into account item and person parameters.

University teachers’ development

The literature on higher-education teachers’ conceptions of and approaches to teaching reveals classifications implying progressive changes from teacher- and content-focused to preferable student-focused conceptions: “from teaching as imparting information to teaching as supporting learning” (McLean and Bullard 2000, p. 83) (also see Barnett and Guzmán-Valenzuela 2017). When teachers systematically hold teacher- or student-focused conceptions of and approaches to teaching, they are considered as being consonant in their teaching. When teachers’ teaching intentions and strategies consist of a varying mix of content- and student-focused conceptions of and approaches to teaching, they are considered to be dissonant in their teaching. Conceptions of and approaches to teaching can be placed on a continuum with the consonant teacher- and student-focused ones on opposite ends. Dissonant teaching profiles are located in between the consonant extremes of the continuum (Postareff et al. 2008; Stes and Van Petegem 2014).

Teacher development theories focus on how teachers change their teaching and develop from a teacher-focused to a student-focused teaching approach. According to Åkerlind (2003), conceptions of teaching are arranged in a hierarchy of expanding awareness of the various aspects of teaching. Teachers start with a teacher-focused conception of teaching and progress towards a student-focused one along the following path: a focus on teacher transmission, a teacher–student relations focus, a focus on student engagement and a focus on student learning. The cumulative awareness of teaching develops along this path because “quantitative understandings of learning [accumulation of information; content-focused approach] have been shown to occur without an awareness of the possibility of qualitative forms of learning [learner-focused approach], but not vice versa” (Åkerlind 2003, p. 377).

At the same time, Åkerlind (2003) states that teachers also expand their awareness of their own professional development as teachers, along the path from a focus on self to a focus on the other (i.e. the learning of the student), and from increasing teacher comfort with teaching and teaching practice to a focus on learning outcomes for students. The focus underlying the two first categories is primarily the quality of teaching, but not yet on the quality of learning. This development path resembles the one described by Kugel (1993) who mentions that university teachers develop in three stages: starting with a focus on self, continuing with a focus on the task and then moving to focus on the learning of the student.

The combination of an increasing understanding of teaching with an expanding awareness of professional development implies that, when teachers transition from a content-focused to a student-focused approach (conception of teaching), they have to gain comfort in each stage before focusing on becoming more effective in that stage (understanding of teaching development). This led Åkerlind (2003) to raise “the possibility of a hierarchy of expanding awareness of the combined aspects of teaching and teaching development” (Åkerlind 2003, p. 388). In this hierarchy, an increasing understanding of teaching, from a focus on transmitting content to a focus on student learning, is combined with an increasing understanding of teaching development, from a focus on teacher comfort alternating with a focus on teaching practices before achieving a focus on student learning. The simultaneous advancing awareness of teaching and teaching development evolves along the following path: (1) teacher transmission joined with teacher comfort, (2) teacher–student relations joined with teacher comfort, (3) teacher–student relations joined with teaching practice, (4) student engagement joined with teacher comfort, (5) student engagement joined with teaching practice, (6) student learning joined with teaching practice and (7) student learning joined with student learning (Åkerlind 2003).

The simultaneity of understandings of teaching and teaching development described by Åkerlind (2003) corresponds to the previously-mentioned dissonant teaching profiles in which teachers combine content- and student-focused intentions and strategies to teaching. Before teachers develop a consonant student-centred profile in their teaching, a dissonant one is anticipated. A teacher might have the intention (i.e. change in conception) to teach in a student-focused way, but the actual teaching strategy (change in approach) might take some time to develop (McLean and Bullard 2000; Postareff et al. 2008; Sadler 2012). For example, in Åkerlind’s (2003) stage four, teachers understand teaching as engaging students with the subject by means of active learning activities (a student-focused conception of teaching). At the same time, they focus on making teaching more effortless and gaining confidence in their practice (an awareness that their development of teaching is teacher-focused).

Development of teaching behaviour

Whereas the development of conceptions of and approaches to teaching have received ample attention in the literature, much less is known about how teachers develop their teaching behaviour. Just recently, Stains et al. (2018) observed over 2000 STEM university classes with the Classroom Observation Protocol for Undergraduate STEM (COPUS). Teacher and student behaviours were recorded during 2-min intervals of the lessons. The authors found seven teaching profiles representing a gradual development towards student-focused teaching, from non-active to more-active learning environments. However, the authors recognised that a limitation of their study was that the COPUS instrument records only the frequency and not the quality of teaching behaviour.

In our search for an instrument that focuses on the quality of teaching behaviour, we compared various observation instruments that are mentioned in the literature on classroom observations in secondary, college and higher education (Noben et al. in press). We considered well-known instruments such as the Classroom Assessment Scoring System-Secondary (CLASS-S) (Pianta et al. 2012), the Framework for Teaching (FFT) (Danielson 2013) and the International Comparative Analysis of Learning and Teaching (ICALT) (van de Grift et al. 2014), as well as newly-constructed ones (Torres et al. 2017).

Our comparison revealed that, although the instruments include common concepts, the ICALT observation instrument is the most complete one for measuring observed teaching behaviour. The ICALT distinguishes six domains of teaching behaviour: (1) safe and stimulating learning climate, (2) classroom management, (3) clarity of instruction, (4) intensive and activating teaching, (5) teaching learning strategies and (6) differentiation (Van de Grift 2014). Until currently, this instrument mainly has been used to examine a stage-wise ordering of the teaching skills of secondary education teachers (Van de Grift et al. 2014; Van der Lans et al. 2017, 2018). In these studies, the established hierarchy corresponds to Fuller's (1969) theory of teacher development that categorises teachers' concerns into three stages. In the first development stage, teachers are concerned with themselves (self) and they focus on, for example, establishing interpersonal relationships with their students. In the second stage, teachers are concerned with the task (task), such as the quality of their instructions. Finally, in the third stage, teachers are concerned with their impact on student learning (e.g. recognising the difficulties of individual students in learning) (Van der Lans et al. 2017, 2018).

Fuller's (1969) theory of teacher concerns for studying teacher development is similar in terms of scope and conceptualisations to Kugel's (1993) and Åkerlind's (2003) theories, described earlier, which acknowledge the progressive change from concerns about *self* to *task* to *impact* on student learning (Fig. 1).

As Fig. 1 shows, there is considerable overlap between the frameworks, but Åkerlind's (2003) is the most extensive. Therefore, in this study, we focused on her theory when examining the development of teaching skills. The aim of our study was two-fold: to explore whether a cumulative ordering of teaching skills can be found for a sample of university teachers and whether the ordering corresponds to Åkerlind's (2003) theory of teaching development. Our study was guided by the following research question: "To what extent is there a hierarchy in observed teaching behaviour, and how does this cumulative ordering correspond to the theory of teacher development?" Additionally, we examined the influence of gender on the cumulative ordering because previous research indicates that female teachers are more likely to report student-focused teaching approaches than male teachers (e.g. Stes and Van Petegem 2014).

Although we recognise teacher development as a holistic process involving various competences, for the purpose of this study, teacher development was limited to teachers' teaching behaviour in relation to their conception of teaching and their understanding of teaching development.

Methodology

This study adopted a quantitative design with a convenience sampling method. Data were gathered through direct classroom observations by trained observers.

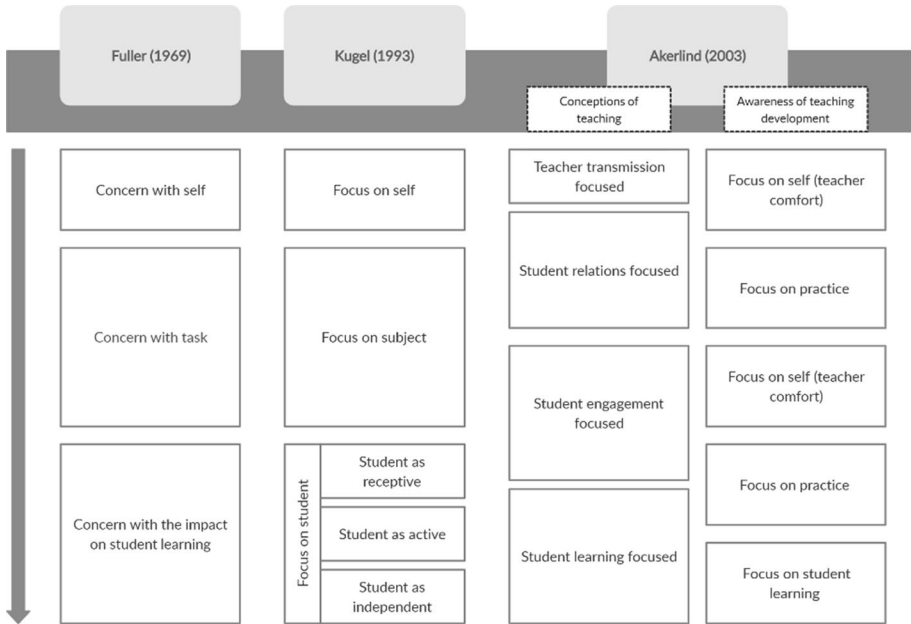


Fig. 1 Teacher development according to Fuller (1969), Kugel (1993) and Åkerlind (2003)

Sample

University teachers were recruited across various faculties by an invitation email. A registry, created to match observers with the interested teachers, contained the following information: a unique number that replaced the teacher’s name, the locations and dates of the teaching moments. After selecting an observation moment in the registry, the observers received the contact details of the teacher and were asked to decline observations of teachers whom they knew personally.

The sample consisted of 211 university teachers whose typical classes were observed once by 24 trained observers at a Dutch research-oriented university. Eight atypical observations were excluded from the sample, including six observations because the lessons were characterised by student presentations with little teacher involvement. Another observation was excluded on the request of the observer. A final observation was discarded because of the particularly challenging nature of the room: the class took place in a movie theatre because of a shortage of classrooms at the university. The final sample of 203 observations included classes from various disciplines, with more undergraduate courses (75%) than Master’s courses (25%). More male ($n = 119$) than female ($n = 84$) teachers were observed and teaching experience ranged from 0 to 44 years ($M = 11.10$, $SD = 9.10$). Class sizes varied between 5 and 220 students ($M = 44.90$, $SD = 45.40$), with 114 small classes (0–30 students) and 87 large classes (more than 30 students). All observations took place between February and July 2017.

Instrument

All observations were undertaken with the ICALT observation instrument (Maulana and Helms-Lorenz 2016; Van de Grift et al. 2014), which includes 32 items measuring observable teaching behaviour indicators, categorised into six teaching domains. The ICALT also includes a three-item student engagement scale, which was used as a criterion variable. This scale specifically focuses on students' psychological and behavioural engagement (Van de Grift et al. 2014). Observers rated all items on a four-point scale, ranging from 1 (mostly weak) to 4 (mostly strong).

Because the ICALT was initially developed for use in primary and secondary educational settings, it was adapted for use in higher education. We modified the formulation of some of the items (face validity at the item level) and three items were added to represent the particular aspects of the higher education teaching context. This resulted in an increase in the number of the items from 32 to 35. The items "The teacher presents societal or research developments of the topic" (Item HE1) and "The teacher compares different theories, perspectives or methods" (Item HE2) were added to the domain of clear and structured instructions. The justification for including the two additional items is that teachers in higher education often teach evolving subjects (i.e. topics covering continuous development of new knowledge) which include a level of uncertainty. Therefore, a certain 'breadth of coverage' is required (Marsh 1982; Healey and Jenkins 2006). A third item "The teacher teaches learners how to reflect" (Item HE3) was added to the domain teaching learning strategies. Teaching students to reflect is widely acknowledged as a goal in higher education. Teachers play a paramount role in supporting students to develop reflection skills (Ryan 2013; Ryan and Ryan 2013).

Data analysis

To examine whether a hierarchical ordering of behavioural domains in teaching behaviour can be found in the higher education context, we applied a Rasch modelling approach. Rasch analysis measures one unidimensional variable or trait (Bond and Fox 2015; Boone et al. 2014). Although we do not refute that teaching behaviour is a multidimensional concept (Feldman 2007; Marsh 2007; Trigwell 2001), in this study, teaching quality was viewed as a unidimensional latent construct consisting of six domains. Focusing on a single aspect of teaching quality (i.e. the latent variable classroom teaching behaviour) can deepen our knowledge of the aspect and guide teachers' professional development in a stage-wise manner (Wright and Stone 1999).

The Rasch measurement model differs from classical test theory (CTT) in various features (Andrich and Marais 2019). Whereas CTT assumes a linear relation between a person's score and the sum of the scores of the individual items, the non-linear Rasch model expresses an estimation of a person's ability (θ parameter) and item difficulty (b parameter) on a single scale. Teachers who are more skillful in teaching behaviour have more likelihood of obtaining a higher score on the observation instrument. Easier items represent behaviour that most teachers are able to demonstrate (Bond and Fox 2015). These simplified principles of the Rasch model allow easy-to-understand feedback to teachers about their level of teaching performance (Van de Grift et al. 2014; Van der Lans et al. 2017). In order to apply the traditional dichotomous Rasch model, recoding of the data is necessary. The values 1 and 2 were recoded to 0 'not observed' and 3 and 4 were recoded to 1

‘observed’ (see Table 4 in the Appendix). Subsequently, 0 means that the observed teachers are still relatively weak in the observed teaching behaviour, while 1 means that the observed teachers are relatively strong in the observed teaching behaviour. We chose to work with the dichotomous Rasch model instead of the polytomous version of the model because the complexity of the latter might unnecessarily complicate feedback to teachers (Van der Lans et al. 2018). To check whether collapsing categories resulted in substantial loss of information, we compared between polytomous and dichotomous models. While the polytomous model explained 54.6% of the total variance, the dichotomous Rasch model (35 items) only explained 46.0% of the total variance. However, both values are still within the 40–50% threshold (Linacre 2019b). Furthermore, the correlation between the polytomous and dichotomous scores was 0.95, suggesting that dichotomisation did not lead to unacceptable loss of information.

We performed the Rasch analysis using WINSTEPS (Linacre 2019a). To examine the model-data fit, the following steps were taken. First, an exploratory factor analysis (EFA) and a confirmatory factor analysis (CFA) were undertaken in *Rstudio version 3.6.0* as a preliminary check on the assumption of unidimensionality. The EFA was conducted using the *psych* package (Revelle 2019) and the CFA was fitted using *lavaan version 0.6-3* (Rosseel 2012), in which the estimation algorithm used was WLSMV and parameterisation was Theta (Muthen and Asparouhov 2002).

Next, item and person fit were examined using outfit statistics. Item fit was examined in terms of Outfit Mean Squares (MNSQ) and ZSTD values. The cut-off criteria for Outfit MNSQ range between 0.5 and 1.5 and ZSTD values fall between -2.0 and $+2.0$, suggesting reasonable fit (Boone et al. 2014). The initial analyses resulted in the exclusion of eight items because of misfit, leaving 26 items to be retained.

A further unidimensionality test involved principal component analysis (PCA) of standardised residuals within WINSTEPS. Afterwards, a Wright map was produced to examine the cumulative ordering of items. We examined differential item functioning (DIF) for gender of the teacher (DIF contrast of >0.64), as well as person- and item-separation to verify the reliability of the model. Person separation reliability refers to how accurately one can differentiate observed teachers according to the measured variable, or “how consistently our estimates of teacher ability match the observed data” (He et al. 2016, p. 389). Item separation reliability refers to how well the model distinguishes between items.

Finally, to assess the predictive validity of the instrument, we estimated the correlation between the student engagement scale and the model’s person measure, the quantitative measure of a teacher’s teaching skills (Boone et al. 2014).

Results

Unidimensionality

Item 1 “The teacher shows respect for the students in his/her behaviour and language” and Item 2 “The teacher maintains a comfortable atmosphere” were excluded from the analysis because of insufficient variance ($M=1.00$, $SD=0.00$). This lack of variation leads to an inability to explain variance, which can cause a model misfit.

Scree plots from exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) allowed verification of the assumption of unidimensionality. The scree plot from EFA showed one dominant factor (Fig. 2).

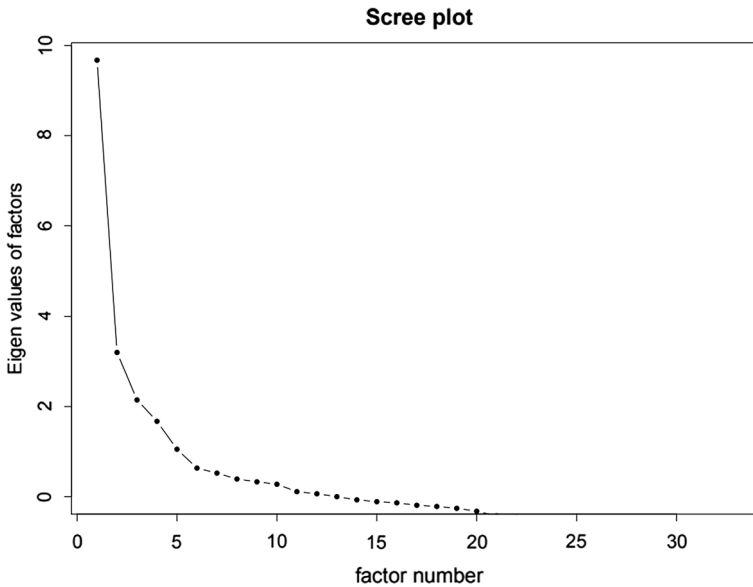


Fig. 2 Scree plot from EFA from 35-item instrument

However, the following robust fit indices of the CFA suggest a relatively poor fit of the data for the one-factor model for teaching behaviour, although the RMSEA value indicated an acceptable fit: $\chi^2(495, N=203)=826.061, p<0.001, SRMR=0.188, RMSEA=0.058$ (CI [0.051, 0.064]), CFI=0.851, TLI=0.841, Gamma Hat=0.92. We used the criteria for acceptable fit of CFI and TLI > 0.90 and RMSEA < 0.08 (Brown 2015; Hu and Bentler 1999). Based on the r-squared item-fit statistics for the individual items, Items 4, 7, 14, HE1 and HE2 (Table 1) were removed from the model. The removal of the items resulted in acceptable fit indices: $\chi^2(350, N=203)=561.710, p<0.001, SRMR=0.168, RMSEA=0.055$ (CI [0.046, 0.063]), CFI=0.914, TLI=0.907, Gamma Hat=0.93. Because χ^2 is sensitive to sample size (Bergh 2015), the decision regarding acceptable model–data fit was primarily based the values of the CFI, TLI, and RSMEA.

The combination of the Outfit MNSQ and ZSTD indices indicated a misfit of two items to the model: Item 5 “The teacher ensures the lesson proceeds in an orderly manner” (Outfit MNSQ = 3.45, ZSTD = 2.3) and Item 22 “The teacher clearly specifies the

Table 1 Item-fit statistic (r^2) for the five misfitting items

Item	Item wording	r^2
	The teacher	
4	... fosters mutual respect	0.003
7	... provides effective classroom management	0.000
14	... teaches in a well-structured manner	0.023
HE1	... presents societal or research developments of the topic	0.010
HE2	... compares different theories/perspectives/methods	0.022

lesson aims at the start of the lesson” (Outfit MNSQ = 1.80, ZSTD = 4.0). To examine whether the fit indices for the misfitting items resulted from an unpredictable score on an item of a specific observation, the impacts of the misfitting entries on the items were examined by considering the absolute ZSTD value of 2.0. The analysis of the 203 responses to Item 5 suggest that a small number of teachers ($n=4$) obtained an unpredictable score (absolute ZSTD value of 2 or higher) on this item, possibly causing an item misfit. Although the scores of these observations on Item 5 were omitted from the analysis (Boone et al. 2014), their removal did not improve the fit (MNSQ = 0.06, ZSTD = -2.0). More teachers ($n=24$) received an unpredictable score on Item 22, but removal of these responses also did not improve the fit (MNSQ = 1.84, ZSTD = 4.1). Consequently, both items were removed from the model.

By removing Items 5 and 22 from the model, Items 9 (MNSQ = 3.07, ZSTD = 2.3) and 23 (MNSQ = 2.27, ZSTD = 2.8) indicated misfit. We further considered the impact of the misfitting entries on these items by considering the ZSTD value of the most misfitting responses. Three observations for Item 9 were replaced by an X in the WINSTEPS control file as recommended by Boone et al. (2014), resulting in adequate fit indices for Item 9 (MNSQ = 1.19, ZSTD = 0.5). When the same procedure was repeated for Item 23, its misfit was resolved (MNSQ = 1.25, ZSTD = 0.80).

The missing entries for Items 9 and 23 resulted in a slight misfit of Item 15 (MNSQ = 1.57, ZSTD = 2.6), but removing one observation for item 15 resulted in satisfactory fit indices for all items (Table 2).

Review of person misfit revealed nine misfitting observations, based on the ZSTD cut-off score. Given that the number of observations to be expected to misfit in this sample ($N=203$) by chance is 10, the total misfitting observations is acceptable (Boone et al. 2014). After closer examination of the misfitting observations, no commonalities were found.

Unidimensional 26-item model

We examined the percentage of total variance explained by the model, as well as item polarity, to verify the unidimensionality of the 26-item model. PCA of the standardised residuals within WINSTEPS indicated that the model accounts for 48.8% of the total variance with an eigenvalue of 24.78, which is slightly below the cutoff of 50%, but within the typical value range of 40–50% (Linacre 2019b). The first contrast accounted for 5.5% of the unexplained variance with an eigenvalue of 2.77 (< 3.0 , Bond and Fox 2015). Finally, we considered the disattenuated correlations of item measures on item clusters. All disattenuated correlations between the item clusters range between 0.8 and 1.00, indicating that the items measure the same latent trait.

The point-measure (PTMEA) correlation indicates whether “the responses to a particular item align with the overall measure” (Bond and Fox 2015, p. 127). Most items showed moderate to high positive correlations (from 0.30 to 0.67), with the exception of Item 3 (PTMEA = 0.29), Item 8 (PTMEA = 0.12), Item 9 (PTMEA = 0.21) and Item 24 (PTMEA = 0.25). The low PTMEA correlations for these items probably indicate that the items are too easy (Items 3, 8 and 9) or too difficult (Item 24) and that “changes in their scores are unlikely to be associated with changes in the overall score” (Ho et al. 2012, p. 317). These four items warrant further analysis.

Table 2 Item measure, standard error, infit and outfit indices, and point-measure correlation for the 26-item model

Item	Measure	SE	Infit		Outfit		PTMEA
			MNSQ	ZSTD	MNSQ	ZSTD	
3	-3.37	0.29	1.34	1.7	1.71	1.3	0.29
6	0.02	0.17	0.97	-0.4	0.88	-0.6	0.57
8	-4.95	0.45	1.21	0.7	2.52	1.7	0.12
9	-4.76	0.43	1.13	0.5	1.31	0.6	0.21
10	-1.51	0.20	0.80	-1.9	0.67	-1.4	0.64
11	-0.25	0.17	0.87	-1.7	0.77	-1.2	0.62
12	-0.62	0.18	0.83	-2.0	0.73	-1.4	0.64
13	-0.10	0.17	1.08	1.0	0.96	-0.1	0.53
15	-0.34	0.17	1.26	3.1	1.35	1.7	0.43
16	-0.90	0.18	1.01	0.2	1.09	0.5	0.55
17	2.24	0.21	1.07	0.6	1.17	0.5	0.39
18	-0.65	0.18	1.02	0.3	1.31	1.4	0.54
19	-0.62	0.18	0.84	-2.0	0.71	-1.5	0.64
20	-0.01	0.17	0.79	-2.9	0.66	-1.9	0.65
21	0.05	0.17	0.76	-3.5	0.62	-2.2	0.67
23	2.10	0.21	1.18	1.5	1.25	0.8	0.35
24	4.02	0.36	1.16	0.6	0.96	0.1	0.25
25	3.39	0.29	0.89	-0.4	0.60	-0.7	0.40
26	4.47	0.42	0.82	-0.4	0.52	-0.7	0.32
27	-0.07	0.17	1.34	4.0	1.33	1.6	0.41
28	0.93	0.18	0.89	-1.5	0.72	-1.2	0.57
29	0.43	0.17	1.10	1.3	0.98	0.0	0.50
30	-0.84	0.18	1.11	1.2	1.04	0.3	0.52
31	-0.13	0.17	0.99	-0.1	0.99	0.0	0.56
32	0.96	0.18	0.90	-1.3	0.71	-1.2	0.57
HE3	0.51	0.17	0.95	-0.6	1.25	1.2	0.55

Person-item map

Figure 3 presents the person and item estimate map (Wright map) of the remaining 26 items. The observation entries are plotted on the left side of the vertical line, while the items are plotted on the right side. Items are presented from more-frequently observed, representing items with lower levels of difficulty (at the bottom of the map), to more-rarely observed, representing items with higher levels of difficulty (at the top of the map). Teachers with a higher measure are those who received a higher score on the instrument (e.g. the observer observed higher quality of teaching behaviour in the observed lesson). Items with a higher measure (e.g. Item 24: measure = 4.02) describe teaching behaviour that was rarely observed by the observers.

Compared with the Wright map of the 35-items version of the instrument (Fig. 4), the relative position of the mean of the observed teachers in the Wright map of the 26-items version (Fig. 3) is highly comparable with that of the mean of the items, which generally suggests high measurement precision for the 26-items version (Boone et al. 2014). However, the revised instrument's map also indicates some degree of measurement imprecision.

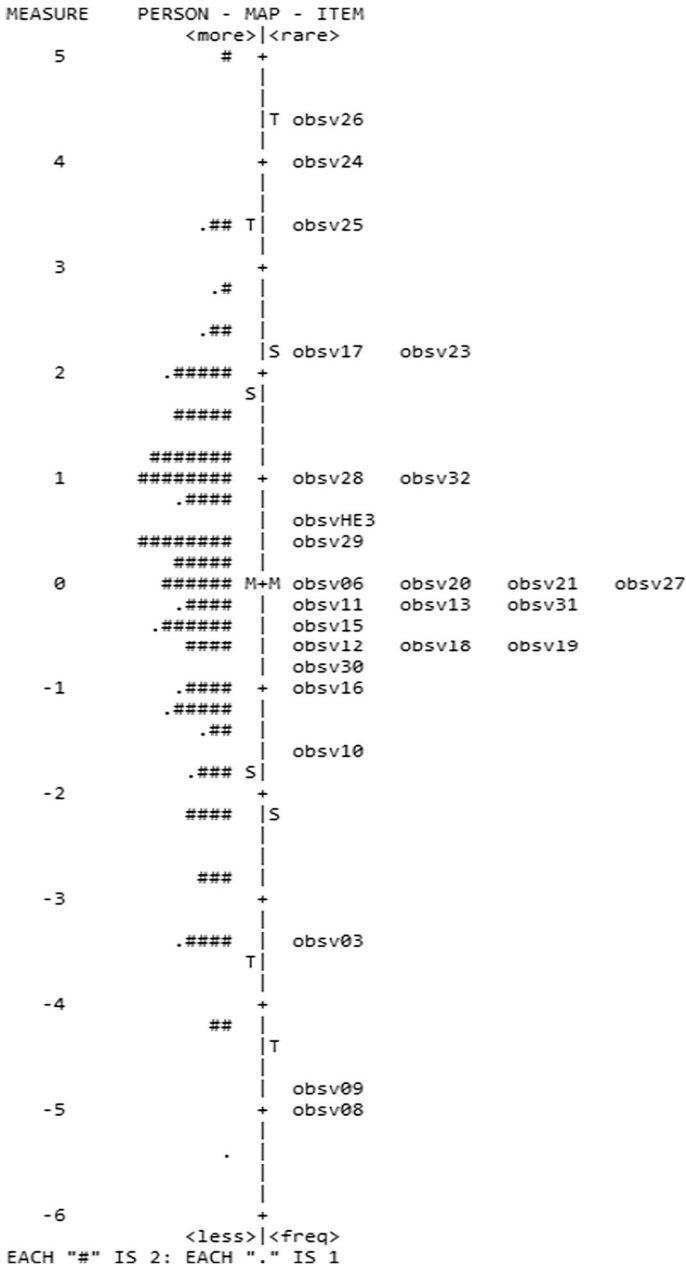


Fig. 3 Wright map 26 items

Several items appear to have the same difficulty level (e.g. Items 12, 18 and 19). Furthermore, the space between Item 3 and 10 suggests the possibility of adding additional items to fill the gap. Nevertheless, the 26-item version overall showed adequate fit to the Rasch model and was retained for further analyses.

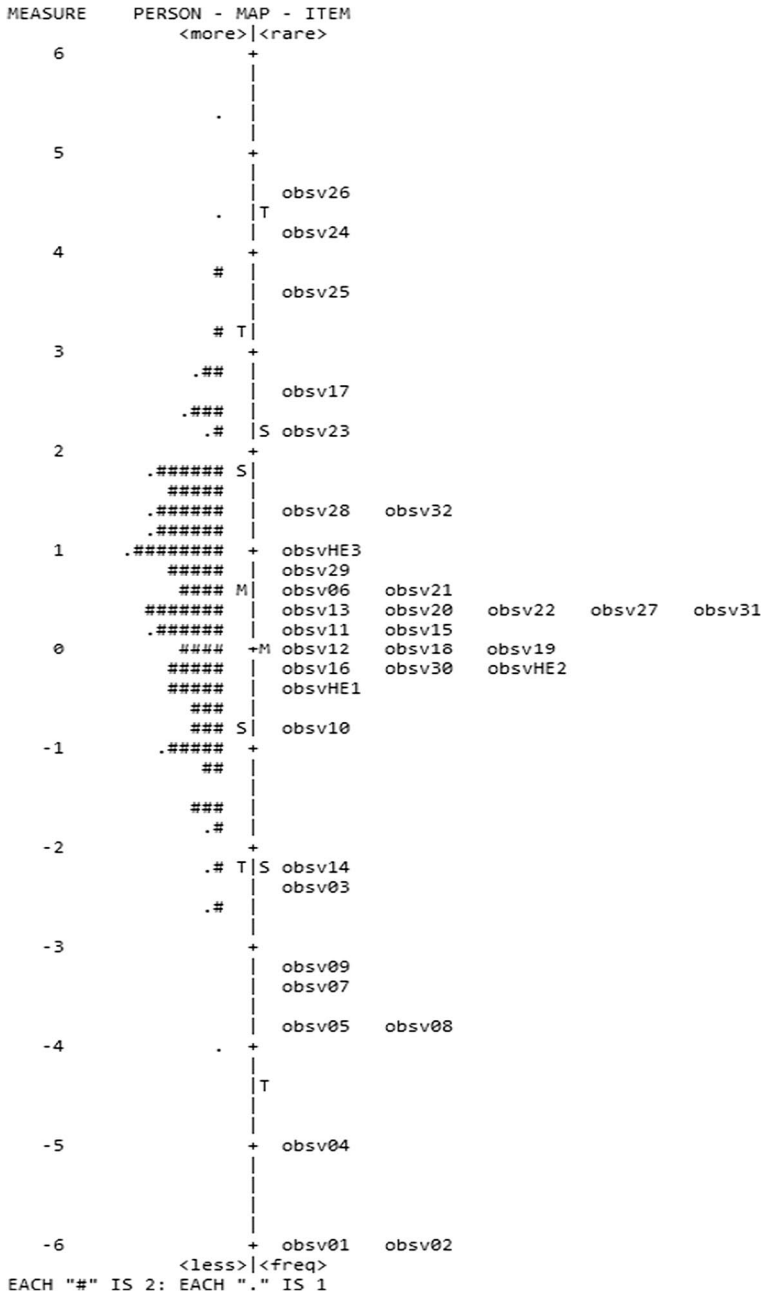


Fig. 4 Wright map 35 items

Differential item functioning (DIF)

In order to examine whether the model was affected by the gender of the teacher (i.e. whether or not the items of the instrument function differently for males and females), a DIF analysis was performed. The Mantel–Haenszel DIF test for dichotomous data indicated that Item 6 (“The teacher monitors to ensure learners carry out activities in the appropriate manner”) possibly showed DIF as a function of gender ($\chi^2=7.92$, $p<0.05$, DIF contrast=11.04!); female teachers seem to demonstrate this behaviour more than male teachers (item group measure -0.61 compared to 0.42) (Boone et al. 2014).

Person- and item-separation reliability

The real person separation index was 2.59 with an associated reliability of 0.87, indicating that the model was able to differentiate between teachers with low and high ability. The real item separation index was 8.55 with an associated reliability of 0.99, indicating that the model can distinguish items that vary in difficulty and that the items set has a high internal consistency (Boone et al. 2014).

Predictive validity

We were able neither to validate the observation instrument with student achievement because of the variety in subjects taught, nor to triangulate the observation data with student evaluation of teachers (SET). Therefore, predictive validity of the instrument was examined by means of the relationship between the teaching skills score with the student engagement scale. There was a strong correlation, meaning that students were more engaged with teachers who obtain higher scores on the instrument, and vice versa ($r=0.59$, $p<0.001$).

Cumulative ordering and teacher development

Table 3 presents the obtained hierarchical ordering of the 26 items according to the b -parameter. A higher b -parameter (i.e. more positive) indicates more complex teaching behaviour. The ordering includes teaching behaviour from all six teaching domains covered by the ICALT. The domains of safe and stimulating learning climate and classroom management were reduced to one and two items, respectively, because of the removal of misfitting items from the model.

To answer the second part of the research question—regarding how the cumulative ordering correspond to the theory of teacher development—the cumulative ordering of the 26 items was considered in relation to Åkerlind’s (2003) theory (Table 3). The ordering largely corresponded to Åkerlind’s (2003) theory of simultaneous advancing of understanding of teaching and teaching development; teachers expanding awareness of teaching can be interpreted in terms of a shift from transmission-focused to more student-focused teaching practice. Throughout this development, teachers alternate between gaining comfort in an aspect of teaching and improving their practice. Eventually they focus on helping students to become independent learners. The stage-wise ordering broadly suggests that

Table 3 The 26-item model compared to Åkerlind's (2003) theory of teacher development

ICALT observation instrument		Teacher development (Åkerlind 2003)	
Item	Item wording	Domain	<i>b</i>
8	The teacher ... uses the time for learning efficiently	Classroom management	-4.95
9	... presents and explains the subject material in a clear manner	Instruction	-4.76
3	... promotes students' self-confidence	Climate	-3.37
10	... gives feedback to learners	Instruction	-1.51
16	... offers activities and methods that stimulate learners to take an active approach	Activating	-0.90
31	... encourages learners to think critically	Learning strategies	-0.84
18	... stimulates learners to think about solutions	Activating	-0.65
12	... during the presentation stage checks whether learners have understood the subject material	Instruction	-0.62
		Conceptions of teaching	Transmission focused
			Awareness of teaching development
			Teacher comfort focused
			Teacher–student relations focused
			Teaching practice focused

Table 3 (continued)

ICALT observation instrument		Teacher development (Åkerlind 2003)			
Item	Item wording	Domain	<i>b</i>	Awareness of teaching development	
				Conceptions of teaching	
				Student engagement focused	
				Teacher comfort focused	
19	... asks questions which stimulate learners to reflect	Activating	-0.62	Student engagement focused	Teacher comfort focused
15	.. gives a clear explanation of how to use didactic aids and how to carry out assignments	Instruction	-0.34		
11	... engages all learners in the lesson	Instruction	-0.25		
30	... stimulates the application of what has been learned	Learning strategies	-0.13		
13	... encourages learners to do their best	Instruction	-0.10		
27	... teaches students how to simplify complex problems	Learning strategies	-0.07		Teaching practice focused
20	... lets learners think aloud	Activating	-0.01		
6	... monitors to ensure learners carry out activities in the appropriate manner	Classroom management	0.02		
21	... gives interactive instructions	Activating	0.05		
29	... teaches learners to check solutions	Learning strategies	0.43	Student learning focused	
HE3	... teaches learners how to reflect	Learning strategies	0.51		
28	... stimulates the use of control activities	Learning strategies	0.93		Student learning focused
32	... asks learners to reflect on practical strategies	Learning strategies	0.96		
17	... stimulates the building of self-confidence in weaker learners	Activating	2.24		
23	... evaluates whether the lesson aims have been reached	Differentiation	2.10		
25	... adjust instructions to relevant inter-learner differences	Differentiation	3.39		
24	... offers weaker learners extra study and instruction time	Differentiation	4.02		
26	... adjust the processing of subject material to relevant inter-learner differences	Differentiation	4.47		

university teachers first develop their skills in the domains of clarity of instruction and intensive and activating teaching before they develop more-complex skills in the domains of teaching learning strategies and differentiation.

The relationship between the items and the categories is based on Åkerlind's (2003) description of the categories. For example, Item 12 "The teacher checks during the presentation stage whether learners have understood the subject material" falls within the category teacher–student relations focused, while Item 19 "The teacher asks questions which stimulate learners to reflect" is categorised as student engagement focused. In the first category, whose aim is to establish good relations with students and teaching behaviour, the focus primarily is what the teachers does within the teacher–student relationship. In the latter category, the aim is to engage students with the topic. Further, Item HE3 "The teacher teaches learners how to reflect" is the last item within the category of teaching practice focused and Item 28 "The teacher stimulates the use of control activities" is the first item in the category of student learning focused. This distinction is based on the verbs in the items; 'teaches' implies a stronger focus on the teacher, while 'stimulates' focuses on students' learning outcomes.

The cumulative ordering of items mostly corresponds to Åkerlind's (2003) theory. However, the position of some items in the model do not match their corresponding categories in Table 3. For example, Item 31 "The teacher encourages learners to think critically" is categorised as a teacher–student relations understanding of teaching combined with a teaching practice focused awareness of teaching development. A better fit for the item would be a student learning focused understanding of teaching and teaching development.

Conclusions and discussion

In this study, our aim was to examine whether a hierarchy in observed teaching behaviour is visible in the context of university teaching. Rasch analysis of observation data confirmed that 26 out of the 35 items included in the modified ICALT showed a cumulative ordering of teaching skills. There is an indication that university teachers develop their teaching by gradually transitioning from focusing on transmitting the content to focusing on the learning of the student. With regards to Åkerlind's (2003) theory of teachers' expanding awareness of their own teaching development—which suggests that teachers alternate between gaining comfort (focus on self) and advancing their knowledge and practice in a specific teaching domain (focus on task) before eventually concentrating on the outcomes of students (focus on the student)—the stage-wise development of teaching practices found in this study seems to suggest that, while transitioning from teacher- to student-focused teaching, teachers also progress from basic to complex teaching skills.

The hierarchical ordering in our sample of university teachers slightly differs from the secondary-education context in which the ICALT observation instrument has extensively been used. Nevertheless, a similar pattern can be identified in terms of the progressive focus from *self* to impact on *student* (Van der Lans et al. 2017). The items that represent less-complex teaching skills (e.g. Items 8, 9, 3, 10) are found in both studies on the easy end of the measurement scale, whereas the most complex teaching skills (e.g. Items 17, 23, 25, 24, 26) are on the difficult end of the scale. This suggests that teachers in higher education, like their colleagues in secondary education, largely seem to develop their teaching

skills by starting with a focus on their role as a teacher but eventually focus on the learning of the student.

A few items in this study are classified as easier than in the study focusing on secondary-education teachers. The most distinct ones are Item 16 “The teacher offers activities and methods that stimulate learners to take an active approach”, Item 19 “The teacher asks questions which stimulate learners to reflect” and Item 31 “The teacher encourages learners to think critically”. A possible explanation for the position of Items 19 and 31 in this study’s model could be the distinctiveness of the corresponding behaviour, reflection and critical thinking, in educational practices in a research-oriented university. The position of Item 16 might be explained by the university’s long-term emphasis on active learning in combination with teachers’ thinking. By only integrating activities in their teaching, teachers could view active learning as involving students in activity with the aim of recalling information instead of engaging students in interactive and collaborative work that generates understanding beyond what has been presented (Andrews et al. 2019). On the other hand, Item 6 “The teacher monitors to ensure learners carry out activities in the appropriate manner” appears to be more difficult for teachers in higher education than for those in secondary education. This could be caused by a view of teaching in higher education in which students are seen as adults who are responsible for their own learning and have a minimal need for teacher involvement.

Our 26-item model mostly excludes items in the domains of safe and stimulating classroom climate and classroom management. Although we consider both domains important aspects of teaching quality (Hativa et al. 2001; Feldman 2007), the characteristics of students in higher education—such as a longer learning history in formal education—could reduce the need for measuring multiple items within these domains, which consequently results in a shorter observation instrument. Although Item 22 is known to be misfitting in the Rasch model in secondary education (Van de Grift et al. 2014; Van der Lans et al. 2017), it is somewhat surprising that the item also misfitted in our model because the higher-education literature emphasises the importance of integrating learning outcomes in teaching practices (Biggs and Tang 2011). Reasons for this remain inconclusive.

While this study provides a general snapshot of university teachers’ teaching behaviour development, we acknowledge the influence on teacher development of context-related aspects, such as the characteristics of the discipline (Lindblom-Ylänne et al. 2006; Stes et al. 2010), the familiarity of the teaching context (Lindblom-Ylänne et al. 2006) and other factors such as time-availability and the number of students (Uiboleht et al. 2016). Furthermore, development is person-dependent. In order for growth to happen, teachers must allow themselves to be open and vulnerable to feedback, as well as to be intrinsically motivated (Postareff et al. 2008; Kelchtermans 2009). The stage-wise development indication described in this study is not devalued by the notion that teacher development is context- and person-dependent. The reason for this is that an individual teacher might progress at a different pace or cease to develop, and consequently stay dissonant in his/her teaching, but the development will occur along the path from teacher- to student-focused teaching. Development of teaching occurs in a hierarchy of inclusiveness in which increasingly sophisticated views of teaching (student-focused) include less sophisticated ones (content-focused), but not vice versa (Åkerlind 2003).

Limitations

While we have used cross-sectional data to study development of teaching behaviour, we recognise that the described cumulative development of teaching needs to be verified at an individual level over time to be able to truly speak of development (longitudinal design). Hence, results of this study should be interpreted as indications of potential development of teaching behaviour. Another note of caution is because of the potential sample bias and sample size. Our sample consists of teachers who volunteered for the observations and who might be relatively more confident and student-focused than the average university teacher. Further research is necessary to verify whether the cumulative ordering can be found in samples that are more representative and from different higher-education institutions. Potential DIF was detected for Item 6 when the effect of gender on the model was examined. Because the number of observations per group (male–female ratio 119–84) was less than the desirable minimum of 300 per group in dichotomous tests (Scott et al. 2009), further examination of the items in relation to gender is required. Finally, although the observers were trained before the observations to ensure consistency of their scores, future studies could examine inter-rater reliability by pairing observers.

Implications for practice

Teachers experience challenges in translating their student-focused conceptions of teaching to their teaching practice, resulting in dissonant teaching profiles (Sadler 2012). The cumulative ordering of teaching skills presented in this study places observed teachers on the measurement scale and identifies which skills they have acquired and still need to acquire. This information can be used to provide tailored feedback to teachers to support their reflection on their teaching practice. The dynamic interaction of reflecting on one's own teaching practice and on one's conceptions of teaching is anticipated to have a positive result on professional learning (Nevgi and Ljöfström 2015). An example of a professional learning activity that scaffolds feedback and fosters reflection could be developmental peer observations of teaching in authentic teaching environments as part of a professional learning community that focuses on developing teaching behaviour. In conclusion, the present study shows that the 26-items version of ICALT is reliable and valid for measuring university teachers' teaching behaviour, which can be used further as a diagnostic tool for university teachers' teaching development over time by supporting them based on their zone of proximal development (Van de Grift et al. 2014).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, 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 licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence 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. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Table 4 Frequency of original ICALT observation scores 1–4 compared with binary recoded scores 0–1

Item	Frequency of original scores				Frequency of binary score	
	1	2	3	4	0	1
1	0	1	57	153	1	210
2	0	2	65	144	2	209
3	0	21	96	94	21	190
4	2	3	108	98	5	206
5	2	8	90	111	10	201
6	71	35	76	29	106	105
7	6	6	106	93	12	199
8	0	7	88	116	7	204
9	6	11	93	101	17	194
10	21	36	93	61	57	154
11	23	73	72	43	96	115
12	28	58	75	50	86	125
13	34	68	83	26	102	109
14	5	22	101	83	27	184
15	51	41	83	36	92	119
HE1	50	21	77	63	71	140
HE2	46	37	72	56	83	128
16	17	55	84	55	72	139
17	142	33	31	5	175	36
18	35	48	93	35	83	128
19	17	66	78	50	83	128
20	62	42	62	45	104	107
21	46	60	69	36	106	105
22	59	48	58	46	107	104
23	129	39	38	5	168	43
24	185	16	10	211	201	10
25	167	27	16	1	194	17
26	177	27	7	0	204	7
27	63	42	78	28	105	106
28	88	51	61	11	139	72
29	78	41	73	19	119	92
30	33	45	102	31	78	133
31	46	52	70	43	98	113
32	87	50	58	16	137	74
HE3	65	57	65	24	122	89

Appendix A

See Table 4.

References

- Åkerlind, G. S. (2003). Growing and developing as a university teacher-variation in meaning. *Studies in Higher Education*, 28(4), 375.
- Andrews, T. C., Auerback, A. J. J., & Grant, E. F. (2019). Exploring the relationship between teacher knowledge and active-learning implementation in large college biology courses. *CBE—Life Sciences Education*, 18(48), 1–17.
- Andrich, D., & Marais, I. (2019). *A course in Rasch measurement theory*. Singapore: Springer.
- Barnett, R., & Guzmán-Valenzuela, C. (2017). Sighting horizons of teaching in higher education. *Higher Education*, 73(1), 113–126.
- Bergh, D. (2015). Chi squared test of fit and sample size: A comparison between a random sample approach and a Chi square value adjustment period. *Journal of Applied Measurement*, 16(2), 204–217.
- Biggs, J., & Tang, C. (2011). *Teaching for quality learning at university* (4th ed.). Open University Press.
- Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York: Routledge/Taylor & Francis Group.
- Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Dordrecht: Springer.
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York: The Guilford Press.
- Cavanagh, R. F., & Waugh, R. F. (2011). *Applications of Rasch measurement in learning environments research*. Rotterdam: Sense Publishers.
- Danielson, C. (2013). *The framework for teaching: Evaluation instrument*. Princeton, NJ: The Danielson Group.
- Feldman, K. A. (2007). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 93–144). Dordrecht: Springer.
- Fuller, F. F. (1969). Concerns of teachers: A developmental conceptualization. *American Educational Research Journal*, 6, 207–226.
- Hativa, N., Barak, R., & Simhi, E. (2001). University teachers: Knowledge and beliefs regarding effective teaching dimensions and strategies. *The Journal of Higher Education*, 72(6), 699–729.
- He, P., Liu, X., Zhen, C., & Jia, M. (2016). Using Rasch measurement to validate an instrument for measuring the quality of classroom teaching in secondary chemistry lessons. *Chemistry Education Research and Practice*, 2, 381–393.
- Healey, M., & Jenkins, A. (2006). Strengthening the teaching–research linkage in undergraduate courses and programs. *New Directions for Teaching and Learning*, 107, 43–53.
- Ho, C. M., Leung, A. W. C., Mok, M. M., & Cheung, P. T. M. (2012). Informing learning and teaching using feedback from assessment data: Hong Kong teachers’ attitudes towards Rasch measurement. In M. Mok (Ed.), *Self-directed learning oriented assessment in the Asia-Pacific. Education in the Asia-Pacific region: Issues, concerns and prospects* (pp. 311–334). Dordrecht: Springer.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1), 1–55.
- Kelchtermans, G. (2009). Who I am in how I teach is the message: Self-understanding, vulnerability and reflection. *Teachers and Teaching: Theory and practice*, 15(2), 257–272.
- Kugel, P. (1993). How professors develop as teachers. *Studies in Higher Education*, 18(3), 315–328.
- Linacre, J. M. (2019a). *Winsteps® Rasch measurement computer program*. Beaverton, OR: Winsteps.com.
- Linacre, J. M. (2019b). *Winsteps® Rasch measurement computer program user’s guide*. Beaverton, OR: Winsteps.com.
- Lindblom-Ylänne, S., Trigwell, K., Nevgi, A., & Ashwin, P. (2006). How approaches to teaching are affected by discipline and teaching context. *Studies in Higher Education*, 31(3), 285–298.
- Marsh, H. W. (1982). SEEQ: A reliable, valid, and useful instrument for collecting students’ evaluations of university teaching. *British Journal of Educational Psychology*, 52(1), 77–95.
- Marsh, H. W. (2007). Do university teachers become more effective with experience? A multilevel growth model of students’ evaluations of teaching over 13 years. *Journal of Educational Psychology*, 99(4), 775–790.
- Maulana, R., & Helms-Lorenz, M. (2016). Observations and student perceptions of pre-service teachers’ teaching behavior quality: Construct representation and predictive quality. *Learning Environments Research*, 19, 335–357.
- McLean, M., & Bullard, J. E. (2000). Becoming a university teacher: Evidence from teaching portfolios (how academics learn to teach). *Teacher Development*, 4(1), 79–101.
- Muthen, B., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in mplus*. Los Angeles: University of California.

- Nevgi, A., & Löfström, E. (2015). The development of academics' teacher identity: Enhancing reflection and task perception through a university teacher development programme. *Studies in Educational Evaluation*, 46, 53–60.
- Noben, I., Deinum, J. F., & Hofman, W. H. A. (in press). Quality of teaching in higher education: Reviewing teaching behaviour through classroom observations. *International Journal of Academic Development*.
- Pianta, R. C., Hamre, B. K., & Mintz, S. L. (2012). *The CLASS-secondary manual*. Charlottesville, VA: University of Virginia.
- Postareff, L., Katajaviuri, N., Lindblom-Ylänne, S., & Trigwell, K. (2008). Consonance and dissonance in descriptions of teaching of university teachers. *Studies in Higher Education*, 33(1), 49–61.
- Postareff, L., Mattsson, M., & Parpala, A. (2018). The effect of perceptions of the teaching-learning environment on the variation in approaches to learning-between-student differences and within-student variation. *Learning and Individual Differences*, 68, 96–107.
- Revelle, W. (2019). *psych: Procedures for psychological, psychometric, and personality research* (R package version 1.9.12). Northwestern University, Evanston, Illinois. <https://CRAN.R-project.org/package=psych>.
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36.
- Ryan, M. (2013). The pedagogical balancing act: Teaching reflection in higher education. *Teaching in Higher Education*, 18(2), 144–155.
- Ryan, Ma., & Ryan, Mi. (2013). Theorising a model for teaching and assessing reflective learning in higher education. *Higher Education Research & Development*, 32(2), 244–257.
- Sadler, R. (2012). The challenges for new academics to adopting student-centred approaches to teaching. *Studies in Higher Education*, 37(6), 731–745.
- Saroyan, A., & Trigwell, K. (2015). Higher education teachers' professional learning: Process and outcome. *Studies in Educational Evaluation*, 46, 92–101.
- Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., et al. (2009). A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. *Journal of Clinical Epidemiology*, 62, 288–295.
- Stains, M., Harshman, J., Barker, M. K., Chasteen, S. V., Cole, R., DeChenne-Peters, S., et al. (2018). Anatomy of STEM teaching in North American universities. *Science*, 359(6383), 1468–1470.
- Stes, A., Coertjens, L., & Van Petegem, P. (2010). Instructional development for teachers in higher education: Impact on teaching approach. *Higher Education*, 60(2), 187–204.
- Stes, A., & Van Petegem, P. (2014). Profiling approaches to teaching in higher education: A cluster-analytic study. *Studies in Higher Education*, 39(4), 644–658.
- Torres, A. C., Lopes, A., Valente, J. M. S., & Mouraz, A. (2017). What catches the eye in class observation? Observers' perspectives in a multidisciplinary peer observation of teaching program. *Teaching in Higher Education*, 22(7), 822–838.
- Trigwell, K. (2001). Judging university teaching. *International Journal for Academic Development*, 6(1), 65–73.
- Uiboleht, K., Karm, M., & Postareff, L. (2016). How do university teachers combine different approaches to teaching in a specific course? A qualitative multi-case study. *Teaching in Higher Education*, 21(7), 854–869.
- van de Grift, W. (2014). Measuring teaching quality in several European countries. *School Effectiveness and School Improvement*, 25(3), 295–311.
- 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.
- van der Lans, R. M., van de Grift, W. J. C. M., & van Veen, K. (2017). Individual differences in teacher development: An exploration of the applicability of a stage model to assess individual teachers. *Learning and Individual Differences*, 58, 46–55.
- van der Lans, R. M., van de Grift, W. J. C. M., & van Veen, K. (2018). Developing an instrument for teacher feedback: Using the Rasch model to explore teachers' development of effective teaching strategies and behaviors. *Journal of Experimental Education*, 86(2), 247–264.
- Webster-Wright, A. (2009). Reframing professional development through understanding authentic professional learning. *Review of Educational Research*, 79(2), 702–739.
- Wright, B., & Stone, M. (1999). *Measurement essentials* (2nd ed.). Wilmington, DE: Wide Range Inc.