University of Groningen

Development and evaluation of a questionnaire measuring pre-service teachers' teaching behaviour

Maulana, Ridwan; Helms-Lorenz, Michelle; van de Grift, Wim

Published in:
School Effectiveness and School Improvement

DOI:
[10.1080/09243453.2014.939198](https://doi.org/10.1080/09243453.2014.939198)

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

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

Publication date:
2015

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

Citation for published version (APA):

Maulana, R., Helms-Lorenz, M., & van de Grift, W. (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>

Copyright

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

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

Take-down policy

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

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

Development and evaluation of a questionnaire measuring pre-service teachers' teaching behaviour: a Rasch modelling approach

Ridwan Maulana*, Michelle Helms-Lorenz and Wim van de Grift

Department of Teacher Education, University of Groningen, Groningen, The Netherlands

(Received 10 April 2013; final version received 4 June 2014)

The present study examines the development of a measure tapping students' perceptions of (pre-service) teachers' teaching behaviour to explore the practical value of such a measure in teacher education and teacher professional development programs. From a sample of 1,635 students of 91 pre-service teachers teaching in secondary education in The Netherlands, random subsamples of 809 students of 45 teachers and of 826 students of 46 teachers were used for analyses. Classical test analyses were used as a preliminary approach prior to utilizing Rasch modelling to the data. Additionally, multilevel analyses were used to examine the predictive validity of the measure on student academic engagement as an external criterion. Results revealed that a shortened and representative measure of teachers' behaviour meets the requirements of the Rasch model sufficiently. In addition, the predictive quality of the shortened measure was confirmed. Implications of findings for research and educational practices were discussed.

Keywords: teachers' behaviour; students' perceptions; Rasch analysis; secondary education

Introduction

Teaching behaviour can be conceptualized as teachers' behaviour affecting pupils' learning and outcomes (Creemers, 1994; Sammons, Hillman, & Mortimore, 1995). The behaviour involves personal and professional characteristics, pedagogical knowledge, classroom climate and management, as well as teacher–student relationships (Danielson, 2013; Pianta & Hamre, 2009; Van de Grift, 2007). Empirical research suggests that teachers' behaviour is an important predictor of students' learning and outcomes including academic engagement, motivation, and achievement (Creemers, 1994; Maulana, Opdenakker, Stroet, & Bosker, 2012, 2013; Opdenakker, Maulana, & Den Brok, 2012; Sammons et al., 1995; Scheerens, 1992; Van de Grift, 2007). Supportive and productive teaching behaviour is recognized in the educational effectiveness literature as an effectiveness-enhancing factor (Creemers, 1994; Kyriakides, Creemers, & Antoniou, 2009; Opdenakker et al., 2012). Teaching behaviour seems to be a central factor determining the success of students and, therefore, plays a significant role in education.

However, previous classroom observation studies have shown that teachers demonstrate great variation in effective teaching behaviour in the classroom (Van de Grift, 2007; Van de Grift & Helms-Lorenz, 2012). These cross-sectional studies reveal that effective teaching behaviour is nonlinearly related to years of experience on the job. Particularly, teaching behaviour of pre-service teachers was found to be relatively low, followed by that of senior teachers (> 20 years of teaching experience). The peak of the teaching

*Corresponding author. Email: r.maulana@rug.nl

performance is evident in a group of teachers with 15 to 20 years of teaching experience. Based on these findings, a clear pattern arises that many years of experience seem to be necessary to reveal average high effective teaching behaviour. Countries, such as The Netherlands, with high retirement rates, face the problem of inexperienced teachers filling this gap. This means that pupils will be exposed to less effective teachers for a long period of time, with cumulative negative achievement effects. To combat this, we should search for effective way(s) to bring pre-service- and beginning teachers' performance up to the level of experienced teachers in a shorter period of time. Undoubtedly, it would be beneficial for educational outcomes to have teachers with less experience displaying effective teaching behaviour similar to the level of experienced teachers.

Pre-service teachers obtain their training to become a professional teacher in the teacher education institution and in the school in which they practise their skills. One of the problems that teacher education faces is the development of guidelines for evidence-based support to promote (amongst other objectives) the improvement of effective teaching behaviour during the first years of teaching. Learning to teach effectively requires a great deal of effort and stamina, not unlike requiring any other form of expertise. The scientific research community can aid teacher instructors in their support of professional development of (pre-service) teachers more effectively, by adding to the effective-teacher-behaviour body of knowledge and by translating and communicating the findings to teacher educational practice.

One potential way of addressing this is by making use of a valid and reliable instrument functioning as a diagnostic tool for the evaluation of teaching performance, but also serving as a research device in the teacher effectiveness domain. The instrument should cover evidence-based effective teaching behaviour. One instrument that meets this requirement does exist in the form of an observation instrument (Van de Grift, 2007, 2013; Van de Grift, Helms-Lorenz, & Maulana, 2013). Based on this observation instrument, we have a number of reasons to extend the examination of teaching behaviour by collecting data directly from students. Although an observation instrument is an invaluable measure of teaching behaviour, it is costly because at least a few observers and observations per teacher/class are needed to get a representative picture of teaching behaviour. A student instrument offers a promising way of measuring teaching behaviour. This method is cost-effective and gives an additional perspective from an individual student level which should have stronger predictive power of student achievement compared to observations targeted at the class level (De Jong & Westerhof, 2001). However, an instrument tapping similar effective teaching behaviour, based on evidence-based theoretical considerations, from the perspective of students is not present yet.

In the attempt to fill this gap, the present study was designed to develop and examine a user-friendly student survey instrument that can mutually be used together with an already existing observation instrument measuring effective teaching behaviour. Academic engagement will be used as an outcome measure to prove the predictive validity of the teaching behaviour measure.¹ A valid and reliable instrument measuring the teaching behaviour performance is not only useful for monitoring the teaching performance of pre-service teachers but also for maintaining (and improving) the teaching performance of experienced teachers.

Theoretical framework

Evidence-based effective teaching behaviour

Teacher effectiveness research, as well as classroom environments research, recognizes the conceptualization of teaching behaviour in a holistic perspective (Alton-Lee, 2003;

Kyriakides et al., 2009). In order to teach effectively, not only personal and professional characteristics matter, but learning environments, instructions, and interpersonal domains are also important (Beishuizen, Hof, Van Putten, Bouwmeester, & Asscher, 2001; Maulana et al., 2012; Opdenakker et al., 2012; Van de Grift, 2007). Teachers' teaching behaviour affects the process and the outcome of students' learning (Creemers, 1994; Sammons et al., 1995; Seidel & Shavelson, 2007). In this regard, teaching behaviour is defined as effective when it has significant and positive impacts on student learning and outcomes. Based on reviews of evidence-based effective teaching behaviour research (Van de Grift, 2007, 2013), there are at least six observable domains of teaching behaviour that have been proven to have an impact on students' learning and outcomes. These teaching domains are classified as follows: Safe and Stimulating Learning Climate, Efficient Classroom Management, Clarity of Instruction, Activating Learning, Adaptive Teaching, and Teaching Learning Strategies. The theoretical foundation underlying the six domains was synthesized from the teacher effectiveness research (e.g., Creemers, 1994; Sammons et al., 1995; Scheerens, 1992), combined with the literature on learning environments and teacher support (e.g., Maulana et al., 2012, 2013). Comparable lists of domains, elaborated with domains that are less directly observable, are found in Muijs and Reynolds (2001), Scheerens and Bosker (1997), Pianta and Hamre (2009), Danielson (2013), and Ferguson (2012). Particularly, Danielson's and Pianta and Hamre's frameworks have been recognized as protocols most widely used in the field. Therefore, these two frameworks are discussed further in comparison to Van de Grift's framework.

Clustering the dimension of teaching behaviour in domains is practical for didactical and conceptual purposes only. As Danielson (2013) points out, clustering of behaviour into domains, and breaking the domains down into components and elements, does not mean or suggest that these behavioural domains are not interrelated. The disentangling of the behaviour into a structure that coincides with teaching *tasks* provides a roadmap for novices and for curriculum developers. The domains do not tap every relevant aspect of teaching (like setting high standards; Hattie, 2012) but should rather be viewed as themes that fuel all behavioural domains (Danielson, 2013). The simplification of teacher behaviour can help guide teachers develop their teaching behaviour by focusing on the domains they need to pay attention to (zone of proximal development). The domains proposed by the mentioned authors show content overlap and are related, as all the behaviours contribute to one common dimension: teaching quality.

It goes beyond the scope of this paper to compare different behavioural clustering principles put forward by the mentioned contributors to the field, or even to defend a particular framework. However, it is necessary to clarify our framework and to pinpoint our contribution to this field. Our clustering is restricted to observable behaviour during a lesson (e.g., this coincides with Domain 2, the classroom environment, non-instructional practices in the classroom; and Domain 3, the instructional practices in the classroom; Danielson, 2013). Lesson preparation activities (Domain 1 of Danielson) and professional responsibilities (Domain 4 of Danielson) are excluded as these behaviours are not visible during the lesson. Danielson's Domains 2 and 3 coincide with 4 of our domains, with the exception of teaching learning strategies and differentiation (see Appendix 1).

Additionally, all the domains put forward by Pianta and Hamre (2009) show a similar pattern of behavioural clustering to our clustering. Appendix 1 makes it clear that the Van de Grift domain clustering covers all the domains proposed by Danielson (2013) and Pianta and Hamre (2009) in one instrument. Even though it could be argued that some behaviours should be viewed as themes, they have been incorporated into our instrument as they have been shown to be highly related to student outcomes. The Van de Grift

domains have been used by expert observers (Van de Grift, 2007; Van de Grift, Van der Wal, & Torenbeek, 2011). In the current study, these domains are applied to the development of a student's questionnaire.² Combining the perspectives of experts and students should contribute to a better understanding of teacher quality. Both perspectives have their benefits and doubts, but using both measures might reduce the weaknesses of the single perspectives.

Safe and stimulating learning climate

Good learning climates are safe and stimulating for students to learn. Conducive and productive learning climates are associated with teachers' behaviour in terms of creating a relaxing learning atmosphere, showing respect to students, and ensuring that students respect the teacher and their peers, encouraging self-confidence of students, and facilitating good teacher–student (interpersonal) relationships (Cornelius-White, 2007; Hattie & Clinton, 2008; Opdenakker et al., 2012; Smith, Baker, Hattie, & Bond, 2008; Teodorović, 2011; Willms & Somer, 2001).

Efficient classroom management

Classroom management is an important domain of teaching practices. Research has shown that efficient classroom management is an important predictor of students' learning and outcomes (Carnine, Dixon, & Silbert, 1998; Houtveen, Booij, De Jong, & Van de Grift, 1999; Maulana et al., 2012; Scheerens & Bosker, 1997). Several indicators of teachers' behaviour associated with efficient classroom management include ensuring that the lesson begins and ends on time, managing lesson transition efficiently, minimizing time for task-unrelated matters, dealing with students' misbehaviour efficiently, preparing the lesson well, and displaying a good lesson structure (Creemers, 1994; Marzano, 2003; Maulana et al., 2012; Opdenakker & Minnaert, 2011; Wang, Reynolds, & Walberg, 1995; Yair, 2000).

Clarity of instruction

Research originating from the behaviourism, constructivism, and direct instruction paradigms has shown that instructional clarity is related to student learning performances. In the beginning of the lesson, it is important to inform the lesson objective clearly in order to let students know what they are expected to do during the lesson (Hattie & Clinton, 2008; Locke & Latham, 1990; Mortimore, Sammons, Stoll, Lewis, & Ecob, 1988; Smith et al., 2008). Furthermore, providing a clear lesson structure and displaying a good interchange of explanations and lesson presentations, managing independent work, and dividing individual and group work clearly are important indicators of instructional clarity as well (Creemers, 1994; Kindsvatter, Willen, & Ishler, 1988; Mortimore et al., 1988; Rosenshine, 1980). Additionally, checking that students understand the learning material is also important (Hattie & Clinton, 2008; Kame'enui & Carnine, 1998; Pearson & Fielding, 1991; Pearson & Gallagher, 1983; Rosenshine & Meister, 1997; Smith et al., 2008).

Within the direct instruction perspective, it is important that instructions be made "direct" or "explicit". This perspective recognizes that in order to maintain instructional clarity, teachers should facilitate the activation of students' prior knowledge, make the lesson objective clear, check whether the lesson objective is achieved, give instructions

into several phases combined with providing examples and how to do the tasks effectively. Consequently, it is important that teachers evaluate if students understand the task given and whether they do the task as expected. Should this not be the case, offering direct feedback is advantageous in order to keep students on task.

Activating learning

Within a popular conceptualization of teaching for active learning, it is acknowledged that learning tends to be optimized when teachers' behaviour is directed towards the facilitation of active learning. Promoting students to learn actively, intensifying instructions, and avoiding excessive work seats are related to students' learning and outcomes (Anderson, Evertson, & Brophy, 1979; Denham & Lieberman, 1980; Fisher et al., 1980; Hampton & Reiser, 2004; Lang & Kersting, 2007; Rosenshine & Berliner, 1978). Moreover, activating students' prior knowledge, making use of "advance organizers", and making sure that students are aware of the relevance of the lesson content are teachers' behaviours related to students' learning performances (Nunes & Bryant, 1996; Pressley et al., 1992). A more recent study has shown that an activating learning environment is related to the quality of teacher–student and peer interactions (Meeuwisse, Severiens, & Born, 2010). If the quality of teacher–student interactions improve, students' learning and performances tend to improve as well (Maulana et al., 2012; Opdenakker et al., 2012).

Teaching learning strategies

In the cognition and information processing literature, cognitive strategies have been developed in various subject areas that students can use to help perform higher level operations. A cognitive strategy, which is also referred to as a teaching-learning strategy, is "a heuristic that serves to support students, facilitating the development of internal procedures that enable them to perform the higher-level procedures" (Van de Grift, 2007, p. 134). Within this research strand, the knowledge about metacognitive strategies is well known. This knowledge serves as a framework to help students achieve a higher level of learning skills (Carnine et al., 1998). A metacognitive strategy is a heuristic approach facilitating students to develop procedures in order to acquire a higher level of learning skills. One of these strategies is that teachers can help support students' learning via scaffolding. Scaffolding is a form of temporary support provided by teachers (or by peer students) that functions as a bridge between students' existing and desired skills. Research has shown that teachers who display modelling explicitly, deliver scaffolding, and provide corrective feedback contribute significantly to the performance of their students (Good & Brophy, 1989; Hattie & Clinton, 2008; Houtveen & Van de Grift, 2007; Rosenshine & Stevens, 1986; Slavin, 1996; Smith et al., 2008).

Adaptive teaching

Effective teaching requires that teachers recognize the characteristics of the students they teach because students with different characteristics (i.e., high- versus low-ability level) have different learning needs in order to progress in learning. This implies that even within the same learning material, teachers need to adapt their teaching given the various student characteristics in their classes. For example, teachers may exhibit a traditional lecturing approach to a group of high-ability students, but teachers may need to combine lecturing and modelling strategies together, in a slower pace, when teaching a group of

low-ability students to achieve the desired learning outcome. In order for the adaptive teaching to effectively work, teachers should put emphasis on the needs of each individual student when displaying instructional techniques. In the literature of differentiation, the term “differentiated instruction” is recognized. This term refers to “a philosophy of teaching purporting that students learn best when their teachers effectively address variance in students’ readiness levels, interests, and learning profile preferences” (Tomlinson, 2005, p. 263). The main goal of differentiated instruction is to maximize the learning potential of each student (Tomlinson, 2003). Research has shown that differentiated instruction leads to a better learning performance (Koeze, 2007; Reis, McCoach, Little, Muller, & Kaniskan, 2011). Several indicators of teaching behaviour associated with adaptive teaching include devoting extra time and additional instructions, pre-teaching and re-teaching, and implementing various effective teaching methods (Houtveen et al., 1999; Kindsvatter et al., 1988; Lundberg & Linnakylä, 1992; Pearson & Fielding, 1991; Sijstra, 1997).

Measuring teachers’ teaching behaviour

Observation instrument

Teacher’s behaviour can be observed by trained teacher observers or it can be tapped by using questionnaires. Previously, an observation instrument measuring teachers’ teaching behaviour targeted at the classroom level in primary education was developed and validated in several European countries including The Netherlands, Germany, Belgium, and England (Van de Grift, 2007, 2013; Van de Grift et al., 2011). Afterwards, a secondary education version of the observation instrument was developed and validated in The Netherlands (Van de Grift et al., 2013). The development of this observation instrument was based on evidence-based research originating from the educational effectiveness framework reviewed earlier (see the theoretical framework section). Both primary and secondary education versions of the instrument were examined using the Rasch analysis and were shown to be reliable and valid (Van de Grift et al., 2013; Van de Grift et al., 2011). Consequently, both versions of the instrument have been used for research purposes as well as for the evaluation tool of pre-service and in-service teachers’ teaching performance.

Student instrument: perceptions of teacher’s behaviour

While an observation instrument is considered to be the most objective tool for measuring teaching practices and could be used as an important procedure in the teacher training process (Lasagabaster & Sierra, 2011; Worthen, Sanders, & Fitzpatrick, 1997), an instrument tapping student perceptions of teachers’ behaviour contributes to the understanding of teachers’ teaching practices in a complementary way. Most notably, student perceptions are based on their (classroom) experiences over time, not merely from a single or limited number of observations. Hence, what students perceive may actually be more important than what outsiders would observe since student perceptions steer individual learning behaviour, based on own judgements and insights. Whether students will academically engage in and successfully complete the expected activities may, to a great extent, depend on their perceptions of the quality and integrity of teachers’ behaviour (Den Brok, Bergen, Stahl, & Brekelmans, 2004). In line with this argument, past studies have shown that student perceptions are mostly more predictive of student outcomes compared to

observations (De Jong & Westerhof, 2001; Seidel & Shavelson, 2007) and teachers' subjective perceptions of their own teaching behaviour (Maulana, Opdenakker, Den Brok, & Bosker, 2011; Scantlebury, Boone, Kahle, & Fraser, 2001). Additionally, a questionnaire is mostly preferred because it is considered as the most cost-effective research tool compared to observations (Fraser, 1991). A valid and user-friendly instrument tapping effective teaching behaviour from the student perspective is not available, and, therefore, it is our objective to contribute to this caveat.

Reconceptualising effective behaviour from the measurement perspective

When shifting from observing teachers to gaining students' perspectives, it becomes necessary to reconceptualize or unravel the concept of effective teaching from the measurement perspective. Both instruments measure common (group-oriented) and unique (individualized) aspects of teaching behaviour. When observing teachers, the observers rate the average behaviour exhibited by the teacher. For example, let's consider the classroom management domain. The observer takes note of all the managerial behaviours exhibited by the teacher and rates the effectiveness thereof taking the average disruptiveness of the pupils into consideration. This weighing of behaviour results in a rating. The individual student's rating is not weighed. The effective teacher behaviour measured by an observer is an averaged measure, whereas the effective teacher behaviour measured by the student reflects the personal impact of the teacher on his/her as a person. The aggregated student ratings at the classroom level should converge with the observer's rating on the dimensions that require a group approach of the teacher (instruction and safe learning climate).

However, the dimensions that require teachers to adapt their behaviour to specific learning needs will not necessarily converge with observed behaviour ratings. Consider for instance the adaptive teaching domain; the observer is focused on adaptive teaching indicators and will rate these highly when one or two incidences are observed. However, when only two students report high perceived adaptive teaching behaviour, then the aggregated classroom mean rating will be low, whereas the observed rating is high in this example. Effective teaching behaviour in the individualized domain, from the observer's point of view, captures the teacher's potential, and the student's point of view captures the depth, intensity, or quality of the teacher's effective teaching behaviour at the classroom level. Exhibiting effective behaviour to the majority of students reflects higher teaching quality. A reconceptualization of effective teaching behaviour from the theoretical point of view does not seem necessary in our framework as we use the same theoretical basis for our conceptualization of effective teaching behaviour for both instruments.

Rasch analysis

The Rasch model can be considered as a family of item response theory (IRT), which has more stringent assumptions compared to a classical test theory (CTT) family. The usefulness of the IRT model is contingent on the extent to which these assumptions are met (Edelen & Reeve, 2007). In order to satisfy the Rasch model, a set of items in the instrument should meet the following assumptions. First, a set of items must measure one and the same latent skill in a homogeneous way (assumption of unidimensionality). In a technical term, the covariance among the items should be explained by a single underlying dimension. Exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) are a useful preliminary procedure to check this assumption (Cattell, 1966).

Then, the IRT model results can confirm whether or not the unidimensionality assumption produced by the EFA and CFA results is tentative (i.e., very low or very high item slope estimates). Additionally, the unidimensionality assumption can be evaluated according to some grouping variables (Edelen & Reeve, 2007; Van de Grift et al., 2013). The second assumption is that the response to one item may not influence the response to another, except for an influence that can be explained by the latent variable that is the measurement objective of the set of items (assumption of local stochastic independence). Technically, this assumption is subsumed under the unidimensionality assumption and requires that, “given their relationships to the underlying construct being measured, there is no additional systematic covariance among the items” (Edelen & Reeve, 2007, p. 6). The third assumption is that the items in a Rasch scale must have a similar discriminatory power (assumption of parallelism of the item characteristic curves). This means that each item in the Rasch scale should contribute uniquely to the scale. Items with extremely high or low discrimination power indicate a violation of this assumption. A product of an instrument that satisfies the Rasch model is a robust instrument for evaluating the teaching performance with high precision.

Aims and hypotheses

Students’ perceptions of teachers’ teaching behaviour are considered to be an important determinant of students’ outcomes (De Jong & Westerhof 2001; Den Brok et al., 2004; Seidel & Shavelson, 2007). Both student perceptions and observations are important resources in the evaluation of teaching performance of pre-service as well as that of more experienced teachers. However, a student instrument measuring teachers’ teaching behaviour comprehensively, originated from evidence-based research, is scarce. A student instrument can focus on either measuring student perceptions of teacher behaviour towards the class (i.e., “My teacher respects my class”) or on that of teacher behaviour towards individual students (i.e., “My teacher respects me”). In the present study, we opted for the latter focus following the framework that individually oriented teaching behaviour delivers the most important information for teachers (Fuller, 1970). Hence, the main aim of the current study was to examine the quality of a student questionnaire measuring teaching behaviour developed based on evidence-based theoretical considerations similar to the previously developed observation instrument.

Because the student questionnaire was developed based on similar theoretical considerations with the previously developed observation instrument (Van de Grift, 2007; Van de Grift et al., 2013), we hypothesize that the student instrument will show a comparable quality with the observation instrument in terms of the construct quality as indicated by meeting the assumptions of the Rasch model satisfactorily. In addition, based on the literature showing the link between teachers’ teaching behaviour and students’ outcomes (e.g., Beishuzen et al., 2001; Maulana et al., 2012; Opdenakker et al., 2012; Van de Grift, 2007), we hypothesize that students’ perceptions of their teachers’ teaching behaviour will significantly predict their (self-reported) academic engagement, and, thus, this will confirm the predictive quality of the instrument.

Method

Data for this study come from a national, longitudinal, study on the relationship between the development of the quality of teaching behaviour of pre-service teachers and the preparation route of teacher education in Dutch secondary education.³ The data for the

present study stem from the first cohort of the first measurement. For each class participating in the study, students completed the questionnaire measuring their teachers' teaching behaviour.⁴

Data and samples

During the academic year of 2010/2011, a total of 2,117 students of Grade 9 to Grade 13 of 95 pre-service teachers from 41 secondary schools throughout The Netherlands participated in this study. Students were distributed as follows across grades: 30% of the students were in Grade 9, 32% in Grade 10, 21% in Grade 11, 14% in Grade 12, and 3% in Grade 13. All schools were recruited based upon voluntary participation, and all the schools in The Netherlands were approached. The age of the students ranged from 11 to 19 years (mean = median = 16 years, $SD = 1.36$), and 53% of them were girls. Approximately 108 pre-service teachers participated in the first cohort.

In the present study, data from a total of 1,635 students of 91 pre-service teachers were included.⁵ This sample consists of 55% female students, 30% of students from a lower grade (Grade 9), 34% of students from alpha subjects (social sciences), 28% from beta subjects (mathematics and natural sciences), and 38% from gamma subjects (history and language). All students are of Dutch ethnicity. This sample is representative of the Dutch school in terms of region of country, location, school type, and ethnicity. The data were collected in the middle of the school year (in February and March).

Measures

Teachers' teaching behaviour

A student questionnaire was developed to capture student perceptions of their teachers' teaching behaviour. The conceptualization of teaching behaviour contained in the questionnaire is theoretically consistent with the evidence-based teaching behaviour literature reviewed in the theoretical framework section. The items in the questionnaire were developed through the following steps: (a) adjusting items of the previously established observation instrument measuring teaching behaviour (Van de Grift, 2007, 2013; Van de Grift et al., 2013) in terms of language use and formulation in order to meet the comprehension level of secondary school students about the measured latent construct; (b) studying the literature about effective teaching behaviour in order to generate ideas about additional items; and (c) generating additional items based upon the reviewed literature on effective teaching behaviours and the existing scales of the previously established observation instrument.

Both the observation instrument and the student questionnaire are based on the same theoretical framework. To ensure that all items in the student questionnaire were qualitatively comparable with the items and domains of the observation instrument, face validity was performed by the authors together with the original developer of the observation instrument. The observation instrument consists of six domains of teaching behaviour: Safe and Stimulating Learning Climate, Efficient Classroom Management, Clarity of Instruction, Activating Learning, Adaptive Teaching, and Teaching Learning Strategies (Van de Grift, 2007). Based on the steps taken, a set of 64 items was constructed. Consistent with the observation instrument, the 64 items could be grouped into six domains including: (a) Safe and Stimulating Learning Climate (12 items), (b) Efficient Classroom Management (8 items), (c) Clarity of Instruction (11 items), (d) Activating

Learning (17 items), (e) Adaptive Teaching (6 items), and (f) Teaching Learning Strategies (10 items, see Appendix 1). All items were provided on a 4-point Likert scale ranging from 1 (*completely disagree*) to 4 (*completely agree*).

Student academic engagement

Students' self-report of academic engagement measure was used to examine the predictive quality of teaching behaviour. The measure was based on a scale developed by Van de Grift (2007). This measure is theoretically consistent with that of Maulana et al. (2012) with the emphasis on psychological and behavioural engagement. The scale consists of six items provided on a 4-point response, ranging from 1 (*completely not true*) to 4 (*completely true*). Examples of items are "I participate well during the lesson", "I do my best during the lesson", and "I pay attention during the lesson". The internal consistency of the scale is above the satisfactory level (Cronbachs' $\alpha = .88$).

Analytic approach

Out of the total sample for the present study, we divided the data randomly into a development ($n_{students} = 809$, $n_{teachers} = 45$) and validation ($n_{students} = 826$, $n_{teachers} = 46$) sample.

Dimensionality and local independence

To check the assumptions of unidimensionality and local independence, we conducted exploratory factor analysis (EFA) of the 64-item set using the developmental sample. Afterwards, we performed a confirmatory factor analysis (CFA) with the validation sample. All factor analyses were conducted using Mplus (Muthén & Muthén, 1998–2012) and treating all items as dichotomous variables.⁶ The dichotomous variable was derived from the four-response category provided in the instrument (Categories 1 and 2 = 0, Categories 3 and 4 = 1). With the dichotomous data, we performed factor analyses based on the tetrachoric correlation matrix instead of the ϕ correlation matrix (Pearson correlations for dichotomous data). The latter is less suitable for the dichotomous data because it can produce items of similar difficulty to have high loadings on the same factor. The former correlation matrix can mitigate this problem (DeMars, 2010).

To examine the unidimensionality from the EFA, we evaluated the scree plot, eigenvalues, and the magnitude of item loadings on the single-factor solution. Following the EFA, we conducted a single-factor CFA model using the validation sample to confirm the unidimensional construct of teaching behaviour. A model with uncorrelated residuals was tested. If a local dependency among pairs of items is evident, the single-factor model with uncorrelated residuals would not show an acceptable fit. Goodness of model-data fit was evaluated based on five indices: chi-square test, the comparative fit index (CFI), the Tucker-Lewis index (TLI), the root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR) (Schreiber, Nora, Stage, Barlow, & King, 2006).

Fitting the Rasch model and item calibration

Like in many statistical models, estimation of parameters in the IRT model is sensitive to sample sizes. In general, the more complex the IRT model, the larger the sample needs to

be. For the Rasch model, where only one parameter is allowed to vary, that is, the item difficulty level, while the slope and the guessing parameters are fixed to certain values, large sample sizes are not necessary. When the sample size is too large, estimation of the Rasch model parameters can become unstable. Sample sizes as small as 100 are often adequate for estimation of the Rasch model (Edelen & Reeve, 2007). Using the random half of the student data in the developmental sample, we fitted the Rasch model to the data and calibrated the items using IRTPRO (Cai, Thissen, & Du Toit, 2005–2013) and WINMIRA (von Davier, 1994).

First, we examined whether the set of items satisfied the requirements of classical test theory in terms of homogeneity by checking the reliability, mean, and variance of raw scores and the corresponding standard deviations. Next, we examined the fit of the model at the item and the overall model fit (factor) level based on the trace line $S\text{-}\chi^2$ item-fit statistic (Orlando & Thissen, 2003). Additionally, we evaluated the item fit using the Q index and the standardized Q index (zq) proposed by Rost and von Davier (1994). Furthermore, we examined a local dependency among pairs of items by means of $LD\text{-}\chi^2$ statistics based on the local dependence statistics proposed by Chen and Thissen (1997). Misfitting items were deleted in a stepwise manner. Items that were locally dependent on other item(s) were also deleted. We evaluated the change in model fit for each time misfitting items were excluded from the model. A number of items that formed an acceptable Rasch model was retained. In addition to assessing the Rasch model, the two-parameter logistic model allowing the slope parameter for each item to vary, also called the Birnbaum model (Birnbaum, 1968), was examined from the final calibration model to check the local dependence associated with violations in the slope estimate.

To examine the overall Rasch model-data fit, the M_2 statistic was used (Maydeu-Olivares & Joe, 2005). In combination with the M_2 statistic, the RMSEA index was also evaluated. Furthermore, we examined the parallelism of the item characteristic curve of the Rasch model using the log-likelihood ratio test proposed by Andersen (Andersen, 1977) by comparing the item difficulty level between teachers with high and low teaching behaviour performance scores. As a final check on the unidimensionality assumption of the final model, we performed the log-likelihood ratio test by comparing group differences in terms of teaching subject (1 = alpha subjects, 2 = beta subjects, 3 = gamma subjects), student gender (1 = male, 2 = female), and grade level (1 = first and second grades, 2 = third grade and above). The analysis was done using the eRm package from R (Mair, Hatzinger, & Maier, 2013).

Predictive validity of perceptions of teaching behaviour measure

Using data from the developmental and validation sample, we generated the Rasch scores and calculated the correlation between the two Rasch scores (θ 's) as well as the mean difference in the Rasch scores across the raw scores (proportion correct). To examine the predictive quality of the instrument that satisfies the Rasch model, we performed multilevel analyses (with teacher as Level 2 and student as Level 1) in the validation sample. A measure of students' self-report on academic engagement was used as dependent variable. Teaching subject, student gender, and grade level were entered as covariates. All significant results under the 95% confidence interval were retained. The analysis was done using MLwiN (Rasbash, Charlton, Browne, Healy, & Cameron, 2005).

Results

Dimensionality of teaching behaviour

Unidimensionality of a latent construct holds when the dominant factor is strongly evident. This means that the estimation of teaching skill levels is not affected by the presence of minor factors (Embretson & Reise, 2000). Based on the results of the single-factor solution of EFA in the developmental sample using the original 64 items, the scree plot of eigenvalues is strongly suggestive of a single factor, with the first value (32.06) significantly larger than the others (2.84, 2.26, 1.47, 1.37, etc., see Figure 1). Item factor loadings are all positive, ranging from .43 to .88, except for Item 35, which has a factor loading of .16. This result suggests that the unidimensionality of the teaching behaviour construct is deemed reasonable.

If the data fit the single-factor solution considerably well, indices criteria derived from CFA should be met. This includes: The chi-square test should not be significant ($p < .05$), the CFI and TLI value should be $\geq .95$, the RMSEA value should be between $< .06$ and $.08$, and the SRMR value should be $\leq .08$ (Schreiber et al., 2006). Results of the one-factor CFA model in the validation sample of the original 64 items indicate an acceptable model-data fit as well ($\chi^2_{(330)} = 1,047.85$, $p = .00$, CFI = .94, TLI = .99, RMSEA = .05, SRMR = .06). The factor loadings range between .18 and .87. For the chi-square test, the result is significant because the χ^2 value is too large given its corresponding degrees of freedom. Nevertheless, this test is very sensitive to sample sizes, so results tend to be significant when the sample size is large. However, other indices show an acceptable model-data fit. This result confirms that the unidimensionality of teaching behaviour holds in the validation sample.

Based on these results, we determined that the 64-item scale is sufficiently unidimensional for the Rasch analysis. Although it is tempting to improve the model-data fit by removing problematic items at this stage, it is advisable to retain the full set of items for the Rasch analysis because removing items prior to the calibration procedure prohibits gaining information about all items' performance in the entire scale context. IRT applications like the Rasch model are often robust to detecting violations associated with item performances (Edelen & Reeve, 2007).

Figure 1. Scree plot of the eigenvalues of the tetrachoric correlation matrix of the 64 items.

The Rasch model and item calibration

Results of the classical test analysis of the 64-items measure suggest that the set of items is homogeneous in terms of classical test theory (Reliability = .93, Mean_{raw scores} = 43.03, $SD = 16.85$). Furthermore, the instrument has to meet the more stringent assumptions of the Rasch model.

After analysing item fit based on the trace line $S-X^2$ item-fit statistic, with the value of $p < .01$ indicating a poor item fit (Orlando & Thissen, 2003), results show that the majority of the items violate the assumption of the Rasch model. Out of the long form of the 64 items, a total of 40 items are identified to show poor fit and indicate local dependency with other items. Subsequently, stepwise deletions based on poor item-fit statistics and local dependency result in the remaining 24 items that show an acceptable fit to the Rasch model (see Table 1). In the set of 24-items version, all the six teaching behaviour domains are still present, including Safe and Stimulating Learning Climate (3 items), Efficient Classroom Management (3 items), Clarity of Instruction (5 items), Activating Learning (3 items), Adaptive Teaching (7 items), and Teaching Learning Strategies (3 items). The short form of 24 items is shown to be homogeneous (Reliability = .85, Mean_{raw scores} = 17.12, $SD = 6.78$).

Of the 24 items, results of $S-X^2$ item-level diagnostic statistics show that four items show a significant value at $p < .05$. Those items include Item 1 (“My teacher treats me with respect”), Item 2 (“My teacher prepares his/her lesson well”), Item 4 (“My teacher involves me in the lesson”), and Item 7 (“My teacher ensures that I treat others with respect”). However, none of the items has $p < .01$.

Furthermore, Rost and von Davier (1994) suggested that the Q index can vary between 0 (perfect fit, i.e., a Guttman pattern) and 1 (perfect misfit), with 0.5 indicating no relationship between the individual parameter and the reaction to the item. High positive values of the z_q index indicate that the item discrimination is lower than assumed by the Rasch model (under-fit), while high negative values indicate higher discrimination than assumed (over-fit). In our data, the Q-index analysis shows that the value of the Q index ranges between .06 and .16, while the value of z_q ranges between -1.05 and 1.01 . These results indicate that there is no item with a significant deviation for the expected characteristic as predicted by the Rasch model.

To evaluate the local independency assumption, we follow the criteria suggested by Chen and Thissen (1997) as documented in Scientific Software International (SSI, 2011). They suggest that an $LD-X^2$ value > 10 indicates likely local dependency among pairs of items, while the value between 5 and 10 lies in the grey area, suggesting that it may either indicate local dependencies or maybe a result of sparseness in the underlying table of frequencies. The value < 5 indicates likely no local dependencies. An inspection of the $LD-X^2$ statistics shows that none of the items has an excessive value to be considered locally dependent on other items (see Table 2). Two pairs of items have the value of 5.1 (Items 3 and 10) and of 6.3 (Items 10 and 13). Nevertheless, no values exceed 10 for all of the 24 items. This suggests that all 24 items appear to meet the assumption of the local stochastic independence satisfactorily.

As a final check of the local dependency among pairs of items, the slope estimate of each item generated from the Birnbaum model was examined (see Table 1). If the excess dependence is problematic, the items would be expected to have high slope estimates (e.g., > 4) relative to values for the other items (Edelen & Reeve, 2007). Results show that the slope values of all 24 items are not exceedingly high given the full range of item slope values between 1.64 and 3. The average slope value is 2.33. Items 6 and 10 seem to

Table 1. Domain, item-difficulty level (b) and slope (a) parameters of the 24-item scale.

Item content													
Domain	Item	Label	My teacher...	p	b	SE	a	SE	Q index	Zq	S-X ²	p	
Learning climate	1	V21	treats me with respect	.86	-1.40	.20	1.82	.31	.16	.63	26.38	.04	
	2	V68	prepares his/her lesson well	.82	-.90	.18	1.64	.27	.16	.96	31.43	.03	
	3	V26	answers my question	.80	-.68	.18	2.30	.35	.10	-.15	22.20	.22	
	4	V13	involves me in the lesson	.79	-.65	.18	1.97	.30	.13	.43	30.69	.04	
	5	V69	makes clear what I need to study for a test	.79	-.65	.18	2.35	.36	.10	-.11	18.67	.48	
	6	V71	helps me if I do not know, or am unable to do something	.79	-.62	.18	3.37	.55	.06	-.105	21.89	.24	
Learning climate	7	V44	ensures that I treat others with respect	.78	-.56	.17	1.76	.27	.14	.82	31.26	.04	
	8	V79	helps me if I do not understand or am unable to do something	.76	-.36	.17	2.55	.38	.09	-.39	15.81	.54	
Teaching learning strategies	9	V81	explains how I need to do things	.76	-.33	.17	2.63	.40	.09	-.42	24.75	.13	
	10	V78	uses clear examples	.76	-.30	.16	3.36	.53	.07	-.104	25.76	.11	
	11	V45	explains the purpose of the lesson	.74	-.20	.17	2.47	.36	.09	-.28	19.18	.45	
	12	V11	ensures that I do my best	.73	-.09	.16	2.34	.34	.10	-.11	19.67	.42	
	13	V23	explains everything clearly to me	.73	-.09	.16	2.81	.41	.08	-.67	23.23	.23	
	14	V58	pays attention to me	.73	-.09	.16	2.48	.36	.09	-.31	15.76	.67	
	15	V57	applies clear rules	.72	-.02	.16	2.03	.29	.12	.33	19.14	.45	
	16	V39	ensures that I pay attention	.68	.32	.16	2.16	.30	.11	.15	20.88	.23	
	17	V49	connects to what I know or am capable of	.67	.35	.16	2.52	.36	.09	-.36	12.29	.72	
	18	V25	checks whether I have understood the subject matter	.65	.54	.15	1.74	.25	.13	1.01	26.19	.06	
Classroom management	19	V16	ensures that I use my time effectively	.63	.72	.15	1.99	.27	.11	.47	13.43	.64	
	20	V19	makes me feel self-confident with difficult tasks	.62	.78	.15	2.17	.30	.10	.21	11.14	.80	
	21	V53	stimulates my thinking	.60	.90	.15	1.85	.26	.12	.74	14.56	.56	
	22	V54	explains how I should study something	.60	.92	.15	2.58	.36	.08	-.42	20.75	.19	
	23	V74	motivates me	.58	1.05	.15	2.92	.42	.07	-.73	18.69	.23	
	24	V65	knows what I have difficulty with	.54	1.36	.15	2.18	.30	.10	.28	16.71	.34	

Note: p = proportion of correct response, p = significant level, b = item difficulty level parameter from the Rasch model, a = slope parameter from the Birnbaum model (2-parameter logistic model), S-X² = item level diagnostic statistics.

Table 2. Standardized $LD\text{-}\chi^2$ diagnostic statistics for the 24-item scale.

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1	–																						
2	–0.7	–																					
3	–0.3	2.3	–																				
4	1.3	2.9	–0.5	–																			
5	1.6	0.7	1.2	0.8	–																		
6	–0.4	–0.4	–0.7	–0.6	–0.6	–																	
7	2.5	0.9	4.3	3.5	2.0	1.1	–																
8	–0.4	–0.7	–0.4	–0.2	0.5	–0.0	–0.7	–															
9	–0.7	0.2	0.5	0.6	2.6	–0.6	–0.3	–0.0	–														
10	0.5	3.3	5.1	1.0	–0.1	0.1	3.9	–0.6	–0.4	–													
11	–0.5	–0.5	–0.6	–0.4	–0.2	0.7	0.0	–0.6	–0.5	–0.0	–												
12	–0.3	–0.7	–0.4	–0.1	–0.7	–0.6	–0.4	–0.6	2.3	–0.5	–0.7	–											
13	–0.0	–0.3	–0.2	–0.6	5.1	–0.7	0.2	–0.5	0.0	6.3	1.5	–0.1	–										
14	0.1	–0.7	–0.7	–0.7	–0.5	2.3	–0.0	–0.7	0.2	0.1	–0.6	–0.2	–0.6	–									
15	–0.7	0.9	1.7	2.5	2.1	–0.7	–0.7	–0.7	–0.5	–0.6	–0.6	2.3	1.4	–0.6	–								
16	–0.6	–0.7	0.8	–0.7	1.6	–0.7	1.1	–0.0	–0.5	–0.6	–0.6	–0.3	–0.6	–0.3	–0.3	–							
17	–0.5	0.6	–0.7	0.4	0.1	–0.2	4.2	0.8	2.3	3.2	0.4	–0.3	0.9	0.7	–0.2	–0.6	–						
18	3.0	0.9	2.7	0.7	–0.6	–0.3	0.0	–0.3	5.5	0.3	–0.3	–0.7	0.6	0.7	0.5	–0.6	5.5	–					
19	–0.5	0.2	–0.2	0.0	0.1	–0.5	3.4	–0.7	–0.2	0.2	1.6	–0.6	1.2	–0.5	–0.6	–0.7	1.3	–0.5	–				
20	0.5	0.4	–0.7	–0.6	–0.6	0.5	–0.5	–0.4	–0.7	–0.7	0.6	2.9	–0.7	–0.6	–0.6	0.5	–0.2	–0.6	–0.5	–			
21	2.7	–0.5	–0.2	1.1	–0.6	–0.5	–0.5	–0.6	–0.7	–0.2	0.9	–0.4	–0.5	–0.2	–0.4	–0.5	0.1	2.5	–0.5	1.0	–		
22	–0.6	–0.0	0.1	–0.7	–0.5	1.8	–0.6	–0.7	–0.7	0.8	0.8	0.6	–0.4	0.3	–0.7	1.1	–0.5	0.0	3.7	1.5	2.1	–	
23	–0.6	–0.6	0.1	0.4	–0.6	–0.4	1.0	0.0	–0.6	–0.6	–0.7	–0.7	–0.3	0.7	–0.7	1.7	–0.7	0.8	–0.5	0.6	0.3	1.0	–
24	–0.5	–0.3	–0.4	–0.7	–0.2	–0.2	1.4	–0.4	–0.6	–0.3	–0.6	0.3	–0.4	1.3	0.2	–0.5	0.6	3.7	–0.7	–0.4	0.0	2.1	–0.7

Note. Values range < 5 indicates likely no local dependency, 5–10 is in the grey area, may indicate either local dependency or the sparseness of the frequency table, > 10 indicates likely local dependency.

Figure 2. The item characteristic curves of the 24 items.

deviate slightly from the average slope value, which has a value of 3.37 and 3.36, respectively. However, the values of those items are not extremely high (> 4). Taken all these results together, we determine that the 24 items of teaching behaviour satisfy the assumption of local stochastic independence.

Furthermore, we examined the overall model fit. The p value $> .05$ indicates a good fit (Maydeu-Olivares & Joe, 2005). The result of the M_2 statistics indicates a lack of fit ($M_{2(275)} = 327.69, p = .02$). However, the associated RMSEA value is .02, which suggests that the lack of fit indicated by the M_2 statistics is possibly due to a minor amount of “model error”, and, therefore, the overall model-data fit is generally deemed acceptable. Moreover, the result of the Andersen test comparing teachers with high and low teaching behaviour performances is not significant ($\chi^2_{(23)} = 17.98, p = .71$). This suggests that the item characteristic curves of the 24 items are reasonably parallel (see Figure 2). Additionally, the 24 items were checked for the unidimensionality assumption required by the Rasch model. Using the Andersen test to evaluate the log-likelihood ratio between groups, results show that the test based on teaching subject, student gender, and grade level is not significant ($\chi^2_{(46)} = 56.60, p = .14$; $\chi^2_{(23)} = 15.33, p = .88$; $\chi^2_{(23)} = 26.78, p = .27$, respectively). From these results, we can ascertain that the unidimensionality of the teaching behaviour scale is not violated by the teaching subject, the gender of students, and the grade level. Finally, the correlation between the scores generated from the original 64 items and the short 24 items was examined. The result shows that the correlation between the two scores is very high ($r = .97, p < .01$).

Predictive validity of teaching behaviour measure

The validity of the short form (24 items) is assessed further. The correlation between the Rasch scores (θ 's) of developmental and validation samples is very high ($r = .99$). This suggests that the estimate of teaching behaviour performances generated from the developmental and validation samples is highly comparable. The mean difference in the Rasch scores is 0.002. This difference is not significantly different from 0 ($t_{(24)} = .69, p = .50$). This suggests an additional support for the validity of the 24-items measure of teaching behaviour.

To examine the predictive validity of the short-form instrument, the link between the Rasch scores of teaching behaviour and student self-report of academic engagement in the validation sample was evaluated using multilevel modelling (see Table 3). Results show that

Table 3. Results of multilevel analysis to examine the link between teaching behaviour and academic engagement; parameter estimates.

Variable	Academic engagement							
	Model 0		Model 1		Model 2		Model 3	
	Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE
<i>Fixed effects</i>								
Intercept	2.82***	.05	2.60***	.04	2.60***	.04	2.68***	.06
Teaching behaviour			.14***	.01	.14***	.01	.13***	.01
Teaching subject							-.03	.07
Student gender							.08*	.04
Grade level							-.24***	.06
<i>Random effects</i>								
Class level								
Intercept	.09	.02	.03	.01	.03	.01	.02	.01
Intercept x Slope					.000	.003	.000	.002
Slope					.001	.001	.001	.001
Student level								
Residual	.32	.02	.29	.02	.29	.02	.29	.02
Deviance	1486.16		1359.78		1358.89		1316.35	

Note. * $p < .05$, *** $p < .01$.

student perceptions of their teachers' teaching behaviour could positively and significantly predict their self-reported academic engagement ($\beta = .14$, $p < .001$, see Models 1 and 2). About 22% of the variance in students' academic engagement could be explained by their perceptions of teachers' teaching behaviour. The effect of teaching behaviour remains highly significant even after adjusting for some control variables ($\beta = .13$, $p < .001$, see Model 3). These results suggest that student perceptions of teaching behaviour are a valid predictor of their perceived academic engagement. This result also confirms that the short form of the teaching behaviour instrument holds a robust predictive validity.

Moreover, we conducted a generalizability study (Shavelson & Webb, 1991) to determine the explained variance at the student and the teacher level. We used the data of 16 pre-service teachers teaching history, math, Dutch, and English. These teachers were rated by 318 students from four classes of three schools in secondary education. The results reveal that the average amount of explained variance is 32.1% and ranges between 17.0% and 53.5% over the classes (see Table 4). This variance is quite high considering the nature of this student instrument. This means that the differences between student

Table 4. Teacher variance in student questionnaire ($n = 4$ classes, 16 teachers, 318 students).

Class	Class size	Variance components		
		Student	Teacher	% Teacher variance
1	22	15.61	6.99	30.9
2	22	22.20	9.25	29.4
3	20	11.35	13.09	53.5
4	17	21.67	4.43	17.0
Total	81	18.34	8.68	32.1

Table 5. Fixed effect estimates of multilevel analysis ($n = 16$ teachers; 318 students).

	Model 0		Model 1	
Variance to be explained student level	18.34		18.17	
Variance to be explained teacher level	8.68		9.30	
fixed effects				
Intercept	14.49	(3.13)	14.747	(3.41)
Gender of students			.23	(.49)
Age of students			-.44	(.45)
Class of students			.94	(.90)
-2* log likelihood	1861.76		1859.75	

ratings can be explained by perceived differences in teacher behaviour observed by the class as a whole. Approximately two thirds of the differences between ratings can be interpreted, if no rival explanations are evident, as being due to the individual student's perception of the teacher.

Finally, we examined the influence of student characteristics such as gender, age, and class where they belonged on student ratings (see Table 5). The results suggest that these variables have no impact on the student ratings gathered with the student questionnaire.⁷

Conclusions and discussion

In the present study, we developed and examined the quality of a student instrument measuring pre-service teachers' teaching behaviour in secondary education. The investigation of the quality of the instrument was done using Rasch modelling and multilevel analysis. In the following section, important findings of the study will be discussed consistent with the hypotheses formulated earlier. Additionally, limitations of the study, implications, and future perspectives will be elaborated.

In line with the argument of Edelen and Reeve (2007), combining a classical test analysis (EFA and CFA) and Rasch modelling offers a complimentary relationship in the process of developing and evaluating an instrument. Results produced by EFA and CFA provide a good starting point for the evaluation of items prior to the calibration procedure using Rasch modelling. Finally, results produced by Rasch modelling provide rich information to select a shortened, valid, and reliable version of the instrument.

Consistent with our first hypothesis, we found that, in general, the short form of the student instrument showed a satisfactory Rasch model property. Therefore, the instrument is psychometrically comparable with the previously developed observation instrument sharing similar theoretical considerations (Van de Grift et al., 2013). Nevertheless, room for improvement remains possible. Results of the item-fit analysis showed that Items 1, 2, 4, and 7 are significant at the $p < .05$, but those items are not significant at the $p < .01$. Because the Rasch model holds very stringent assumptions and the model rarely fits perfectly, one may expect (slightly) significant values because the model is not perfect (SSI, 2011). Besides, results of the local item dependency analysis showed that none of the items mentioned indicate a violation to the local item dependency assumption. Although two item pairs (Items 3 and 10, Items 10 and 13) have values that fall in the grey area, the values are not extremely high. Most probably, this is caused by the

sparseness in the underlying table of frequencies in the data (the data contain two response categories of 24 items). Given these results, in combination with the verdict that the content of the items is relevant for tapping teaching behaviour, the decision to retain these items seems plausible.

Room for improvement is also indicated by the significant result of the M_2 statistics, which suggests that the overall-model data fit could be improved. Nevertheless, results of the RMSEA suggests that this is a minor issue (deleting some items mentioned earlier could improve the model fit slightly). Because our goal is to construct an instrument that is representative for tapping teaching behaviour (at least 3 items per domain) and at the same time holds the property of the Rasch model sufficiently, the decision to retain the 24-item version is justified. Additionally, we found that the unidimensionality assumption holds for several grouping variables including the teaching subject, the gender of students, and the grade level. This suggests that the instrument can be used widely regardless of these grouping variables.

Even though the Rasch properties were confirmed for the student version, the internal structure of the construct differs from that of the peer observation version (Van de Grift et al., 2013). The results of the Rasch modelling revealed that the patterning of the item difficulty level parameters for pre-service teachers differs from that of item difficulty level parameters yielded from observations of pre-service teachers. One partial explanation could be that observers view different behaviour caused by their presence/intrusion. What is more important, students evaluate personally directed teacher behaviour (weighing whether the observed behaviour is powerful enough to serve personal student aims), whereas observers judge overall teacher behaviour (weighing whether the observed behaviour is powerful enough to serve classroom aims). These personal versus general orientations lead to non-identical structures. The peer data reveal an hierarchical structure of increasing difficulty of behaviours clustered in domains. This clustering of behaviour is less evident in the student version. Additionally, Efficient Classroom Management is not judged by students as being an easy skill as it reveals medium-high difficulty levels. Observers seem to judge this aspect more leniently. The clustering of behaviour into domains serves didactical and feedback purposes. The implication of this finding is that both instruments reveal two different aspects of effective teacher behaviour, an effective classroom approach and an effective individual student approach.

Our second hypothesis was confirmed. We found that student perceptions of teachers' behaviour could significantly predict their academic engagement. Results suggest that the better the teaching behaviour perceived by students, the higher the level of academic engagement tends to be. In line with the existing body of knowledge about the importance of teachers' behaviours for students' learning and outcomes (Beishuizen et al., 2001; De Jong & Westerhof, 2001; Den Brok et al., 2004; Maulana et al., 2011, 2012; Opendakker et al., 2012; Van de Grift, 2007), the present study confirms that perceptions of teachers' teaching behaviour is a significant determinant of students' academic engagement. Additionally, this finding also suggests that the short form of the student instrument has sufficient predictive validity. Importantly, we used the Rasch scores instead of the raw scores of teachers' teaching performance to predict students' academic engagement. Besides that the Rasch scores are a robust estimate of the teachers' teaching performance quality, our finding suggests that the significant predictive validity of teaching behaviour generated from the Rasch scores provides an evidence for another potential advantage of using Rasch scores in predicting various student outcomes.

Finally, it is the global educational effectiveness goal to explore what works why for students and how to improve teaching practices to achieve this. The present study adds to

the knowledge base by offering a novel, user-friendly, and valid instrument measuring student perceptions of teachers' teaching behaviour generated from evidence-based theoretical considerations and analysed by a robust Rasch model, and by confirming that student perceptions of teachers' teaching behaviour is an important predictor of students' academic engagement. This implies that focusing on improving the quality of evidence-based teachers' behaviour already from the pre-service teaching context would potentially contribute to help inexperienced teachers perform to the expected level. Therefore, effective teachers' behaviour and the use of its corresponding valid measure should be implemented in teacher education institutions and schools.

Limitations and future directions

The main purpose of the present study was to investigate the psychometric quality of a student instrument measuring evidence-based teachers' behaviour by utilizing the Rasch model in selecting a shortened and representative scale in the context of pre-service teachers teaching in secondary education. Although our findings suggest that the short, 24-item, version of the student instrument is psychometrically reliable and valid, this is true as far as the context of secondary education pre-service teachers is concerned. Thus, the instrument can potentially be used for research and practices in this particular context. Whether or not the instrument is valid for the whole population of secondary education requires further examination in future research. A comprehensive validation of the shortened scale would be necessary by taking into account a broader sample that includes pre-service and experienced teachers together.

In addition, it would potentially be worthwhile to validate the short-form instrument in an independent sample of students as one may expect that students would respond to a short instrument (24 items) rather differently as opposed to a long one (64 items). Last but not least, because one critical goal of educational effectiveness research is the improvement of students' academic achievement, the absence of achievement data in this study can be seen as another limitation.⁸ Future research would benefit from the inclusion of such data as a dependent factor to be predicted by teachers' behaviour. Because measures of teaching behaviour and student engagement were both collected from students, the interpretation regarding the relation between teaching behaviour and academic engagement is limited to a sense that student *perceptions* of teaching behaviour are positively related to their *perceptions* of academic engagement. Nonetheless, our finding is in line with the research of Van de Grift (2007) and Van de Grift et al. (2013) investigating the same constructs using experts' observations. Regardless of these limitations, our primary goal to develop a representative student instrument for use primarily in teacher education institutions has been accomplished. It is our expectation that this article will draw researchers' attention for a global discussion and future collaboration for the improvement of research and practice in the field.

Implications

Findings of the present study substantiate the relevance of several implications for research and practice. First, results of the Rasch analysis generated from evidence-based research is a robust estimate of teaching performance level. Hence, teacher education institutions and schools might rely on the Rasch scores as a potential gold standard for supporting continuous professional development of teachers. Using the instrument as a diagnostic/feedback tool is recommended. Because the feedback is generated from

individual students' experience with their teachers, teachers are informed regarding their one-to-one interactions with their students. In turn, this will offer precise clues for teachers to which particular students they have to pay more attention to endorse a more personalized teaching having more impacts on student learning. Furthermore, because teaching behaviour is a malleable factor, the opportunity for improving teaching behaviour is possible. This can be done by implementing more precise interventions targeting at the zone of proximal development after the diagnostic attempt using the instrument has been undertaken. Another implication of valid effectiveness measures is that they can be used to evaluate university-based teacher preparation programs (Bell, Maeng, & Binns, 2013; Hildebrand, 2009; Stieger, 2013). Finally, the finding that teacher behaviour could significantly predict student academic engagement supports the current theory about the importance of teacher behaviour for student learning and outcomes (based on students' perceptions). This implies that teachers do matter for students, and effective attempts for improving the teaching performance level are called for.

Acknowledgements

We thank the editor and the two anonymous reviewers for their constructive and useful feedback on the previous draft of this paper. We are also indebted to Rikkert van der Lans for collecting the additional data for this study. Our special thanks also go to the research team, Esther Canrinus and Anna Verkade, for their great assistance during the implementation of the project.

Funding

This study was part of the national research project about the relationship between the development of the quality of teaching behaviour of pre-service teachers and the preparation route of teacher education in Dutch secondary education. This project was financed by the Nederlandse organisatie voor Wetenschappelijk Onderzoek [NWO, 411-09-802]. NWO funds scientific research for Dutch universities and institutes.

Notes

1. For the purpose of predictive validity, the size of the scores on the observation instrument should correlate significantly and strongly with student achievement. Because of the diversity of school types in which teachers worked and the wide range of subjects they taught, it was not possible in this study to validate the instrument with student achievement.
2. The observation instrument was used as the first source to generate items for the student questionnaire in this study. Based on experts' judgements (i.e., face validity), selected and relevant items from the observation instrument were adjusted to meet the comprehension level of secondary school students. All domains in the observation instrument were included in the student questionnaire.
3. In Dutch: OIDS, an acronym for *Opleiding in de School*. OIDS is a longitudinal project that was designed to track various cohorts of pre-service teachers between the 2010 and 2014 academic year in The Netherlands.
4. Students rated teaching behaviour of pre-service teachers and reported their academic engagement. All pre-service teachers taught students as part of practical experience required by their training institute.
5. The total data sample was screened prior to the data analysis. Cases that did not meet criteria to be included in the analysis were excluded. This included classes consisting of fewer than 10 students and cases with a high number of missing cases.
6. The decision to treat the survey responses as dichotomous variables was necessary so the stringent one-parameter Rasch model could be applied to the data. Additionally, our practical intention is to investigate whether or not pre-service teachers displayed the measured teaching behaviour (based on student perceptions). A score of "0" means that the teacher did not display

the measured behaviour, and a score of “1” means that the teacher displayed the measured behaviour.

7. Additional G study using the validation sample was conducted as well. G study, or generalizability theory, is “a statistical theory for evaluating the dependability (‘reliability’) of behavioural measurements” (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Because the sample is limited to only one facet to be included, we performed a single-facet fully crossed design. In this design, the G coefficients are identical to the intra-class correlation that is commonly used to estimate inter-rater reliability (Hoyt & Melby, 1999). Results reveal that the relative and the absolute G coefficients are .92 and .91, respectively. This provides an additional evidence that student ratings are a reliable measure of teaching behaviour.
8. Our analysis of different datasets showed that student perceptions of teaching behaviour are related positively to expert observer ratings of student engagement. This result supported the present study regarding a positive relationship between teaching behaviour and academic engagement.

Notes on contributors

Ridwan Maulana is an assistant professor at the Department of Teacher Education, University of Groningen, The Netherlands. He received his doctorate in educational sciences from the same university. His major research interests involve teaching and teacher education, learning and instruction, as well as statistics and methods associated with the measurements of teaching in the context of teacher education.

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

Wim van de Grift is a full professor and director of the Department of Teacher Education, University of Groningen, The Netherlands. His research interest involves educational effectiveness, professional development of teachers, quantitative methodology, as well as subject-related teaching skills and performances.

References

- Alton-Lee, A. (2003). *Quality teaching for diverse students in schooling: Best evidence synthesis*. Wellington: Ministry of Education, Medium Terms Strategy Policy Division.
- Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, 42, 69–81.
- Anderson, L. M., Everson, C. M., & Brophy, J. E. (1979). An experimental study of effective teaching in first-grade reading groups. *The Elementary School Journal*, 79, 193–223.
- Beishuizen, J. J., Hof, E., Van Putten, C. M., Bouwmeester, S., & Asscher, J. J. (2001). Students and teachers’ cognitions about good teachers. *British Journal of Educational Psychology*, 71, 185–201.
- Bell, R. L., Maeng, J. L., & Binns, I. C. (2013). Learning in context: Technology integration in a teacher preparation program informed by situated learning theory. *Journal of Research in Science Teaching*, 50, 348–379.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’s ability. In F. M. Lord & M. R. Novick. (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.
- Cai, L., Thissen, D., & Du Toit, S. (2005–2013). IRTPRO (Version 2.1) [Computer software]. Lincolnwood, IL: Scientific Software International.
- Carnine, D. W., Dixon, R. C., & Silbert, J. (1998). Effective strategies for teaching Mathematics. In E. J. Kame’enui & D. W. Carnine (Eds.), *Effective teaching strategies that accommodate diverse learners*. Upper Saddle River, NJ: Prentice Hall.
- Cattell, R. B. (1966). The data box: Its ordering of total resources in terms of possible relational systems. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology* (pp. 67–128). Chicago, IL: Rand-McNally.

- Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22, 265–289.
- Cornelius-White, J. (2007). Learner-centered teacher-student relationships are effective: A meta-analysis. *Review of Educational Research*, 77, 113–143.
- Creemers, B. P. M. (1994). *The effective classroom*. London: Cassell.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York, NY: Wiley.
- Danielson, C. (2013). *The framework for teaching: Evaluation instrument*. Princeton, NJ: The Danielson Group.
- De Jong, R., & Westerhof, K. (2001). The quality of student ratings of teacher behaviour. *Learning Environments Research*, 4, 51–85.
- DeMars, C. (2010). *Items response theory: Understanding statistics measurement*. New York, NY: Oxford University Press.
- Den Brok, P., Bergen, T., Stahl, R. J., & Brekelmans, M. (2004). Students' perceptions of teacher control behaviours. *Learning and Instruction*, 14, 425–443.
- Denham, C., & Lieberman, A. (Eds.). (1980). *Time to learn*. Washington, DC: National Institute of Education.
- Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modelling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, 16(1S), 5–18.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
- Ferguson, R. F. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan*, 94(3), 24–28.
- Fisher, C. W., Berliner, D. C., Filby, N. N., Marliave, R. S., Cahen, L. S., & Dishaw, M. M. (1980). Teaching behaviours, academic learning times, and student achievement: An overview. In C. Denham & A. Lieberman (Eds.), *Time to learn* (pp. 7–32). Washington, DC: National Institute of Education.
- Fraser, B. (1991). Two decades of classroom environment research. In H. J. Walberg (Ed.), *Educational environments: Evaluations, antecedents and consequences* (pp. 3–27). Elmsford, NY: Pergamon.
- Fuller, F. (1970). *Personalized education for teachers: One application of the teacher concerns model*. Austin, TX: R & D Center for Teacher Education.
- Good, T. L., & Brophy, J. (1989). Teaching the lesson. In R. E. Slavin (Ed.), *School and classroom organization* (pp. 25–68). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Hampton, S. E., & Reiser, R. A. (2004). Effects of a theory-based feedback and consultation process on instruction and learning in college classrooms. *Research in Higher Education*, 45, 497–527.
- Hattie, J. A. C. (2012). *Visible learning for teachers: Maximizing impact on learning*. Abingdon: Routledge.
- Hattie, J. A. C., & Clinton, J. (2008). Identifying accomplished teachers: A validation study. In L. Ingvarson & J. A. C. Hattie (Eds.), *Assessing teachers for professional certification: The first decade of the National Board for Professional Teaching Standards* (pp. 313–344). Oxford: Elsevier.
- Hildebrand, K. E. (2009). Using curriculum-based measures to assess special education teacher candidates' application of defining and measuring behaviour skills. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, 69, 4290.
- Houtveen, A. A. M., Booij, N., De Jong, R., & Van de Grift, W. J. C. M. (1999). Adaptive instruction and pupil achievement. *School Effectiveness and School Improvement*, 10, 172–192.
- Houtveen, A. A. M., & Van de Grift, W. J. C. M. (2007). Effects of metacognitive strategy instruction and instruction time on reading comprehension. *School Effectiveness and School Improvement*, 18, 173–190.
- Hoyt, W. T., & Melby, J. N. (1999). Dependability of measurement in counseling psychology: An introduction to generalizability theory. *The Counseling Psychologist*, 27, 325–352.
- Lang, J. W. B., & Kersting, M. (2007). Regular feedback from student ratings of instruction: Do college teachers improve their ratings in the long run? *Instructional Science*, 35, 187–205.
- Locke, E. A., & Latham, G. P. (1990). *A theory of goal setting and task performance*. Englewood Cliffs, NJ: Prentice Hall.
- Kame'enui, E. J., & Carnine, D. W. (Eds.). (1998). *Effective teaching strategies that accommodate diverse learners*. Upper Saddle River, NJ: Prentice Hall.

- Kindsvatter, R., Wilen, W., & Ishler, M. (1988). *Dynamics of effective teaching*. New York, NY: Longman.
- Koeze, P. A. (2007). *Differentiated instruction: The effect on student achievement in an elementary school* (Unpublished doctoral dissertation). Eastern Michigan University, Ypsilanti.
- Kyriakides, L., Creemers, B. P. M., & Antoniou, P. (2009). Teacher behaviour and student outcomes: Suggestions for research on teacher training and professional development. *Teaching and Teacher Education*, 25, 12–23.
- Lasagabaster, D., & Sierra, J. M. (2011). Classroom observation: Desirable conditions established by teachers. *European Journal of Teacher Education*, 34, 449–463.
- Lundberg, L., & Linnakylä, P. (1992). *Teaching reading around the world*. The Hague: IEA.
- Mair, P., Hatzinger, R., & Maier, M. J. (2013). *Extended Rasch modeling: The R Package eRm*. Retrieved from <http://cran.r-project.org/web/packages/eRm/vignettes/eRm.pdf>
- Marzano, R. J. (2003). *What works in schools. Translating research into action*. Alexandria, VA: ASCD.
- Maulana, R., Opdenakker, M.-C., Den Brok, P., & Bosker, R. (2011). Teacher-student interpersonal relationships in Indonesia: Profiles and importance to student motivation. *Asia Pacific Journal of Education*, 31, 33–49. doi:10.1080/02188791.2011.544061
- Maulana, R., Opdenakker, M.-C., Stroet, K., & Bosker, R. (2012). Observed lesson structure during the first year of secondary education: Exploration of change and link with academic engagement. *Teaching and Teacher Education*, 28, 835–850.
- Maulana, R., Opdenakker, M.-C., Stroet, K., & Bosker, R. (2013). Changes in teachers' involvement versus rejection and links with academic motivation during the first year of secondary education: A multilevel growth curve analysis. *Journal of Youth and Adolescence*, 42, 1348–1371.
- Maydeu-Olivares, A., & Joe, H. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2nd contingency tables: A unified framework. *Journal of the American Statistical Association*, 100, 1009–1020.
- Meeuwisse, M., Severiens, S. E., & Born, M. P. (2010). Learning environment, interaction, sense of belonging and study success in ethnically diverse student groups. *Research in Higher Education*, 51, 528–545.
- Mortimore, P., Sammons, P., Stoll, L., Lewis, D., & Ecob, R. (1988). *School matters. The junior years*. Somerset: Open books.
- Muijs, D., & Reynolds, D. (2001). *Effective teaching: Evidence and practice*. London: Sage.
- Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide seventh edition*. Los Angeles, CA: Authors.
- Nunes, T., & Bryant, P. (1996). *Children doing mathematics*. Oxford: Blackwell.
- Opdenakker, M.-C., Maulana, R., & Den Brok, P. (2012). Teacher-student interpersonal relationships and academic motivation within one school year: Developmental changes and linkage. *School Effectiveness and School Improvement*, 23, 95–119. doi:10.1080/09243453.2011.619198
- Opdenakker, M.-C., & Minnaert, A. (2011). Relationship between learning environment characteristics and academic engagement. *Psychological Reports*, 109, 259–284. doi:10.2466/09.10.11.PRO.109.4.259-284
- Orlando, M., & Thissen, D. (2003). Further investigation of the performance of $S-X^2$: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, 27, 289–298.
- Pearson, P. D., & Fielding, L. (1991). Comprehension instruction. In R. Barr, M. L. Kamil, P. B. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research* (Vol. 2, pp. 815–860). White Plains, NY: Longman.
- Pearson, P. D., & Gallagher, M. C. (1983). The instruction of reading comprehension. *Contemporary Educational Psychology*, 8, 317–344.
- Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, 38, 109–119.
- Pressley, M., Wood, E., Woloshyn, V. E., Martin, V., King, A., & Menke, D. (1992). Encouraging mindful use of prior knowledge. Attempting to construct explanatory answers facilitates learning. *Educational Psychologist*, 27, 91–109.
- Rasbash, J., Charlton, C., Browne, W. J., Healy, M., & Cameron, B. (2005). *MLwiN (Version 2.0)* [Computer software]. Bristol: University of Bristol.

- Reis, S. M., McCoach, D. B., Little, C. A., Muller, L. M., & Kaniskan, R. B. (2011). The effects of differentiated instruction and enrichment pedagogy on reading achievement in five elementary schools. *American Educational Research Journal*, 48, 462–501.
- Rosenshine, B. (1980). How time is spent in elementary classrooms. In C. Denham & A. Lieberman (Eds.), *Time to learn* (pp. 107–126). Washington, DC: National Institute of Education.
- Rosenshine, B., & Berliner, D. (1978). Academic engaged time. *British Journal of Teacher Education*, 4, 3–16.
- Rosenshine, B. V., & Meister, C. (1997). Cognitive strategy instruction in reading. In S. Stahl & D. A. Hayes (Eds.), *Instructional models in reading* (pp. 85–107). Mahwah, NJ: Erlbaum.
- Rosenshine, B. V., & Stevens, R. (1986). Teaching functions. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 376–390). New York, NY: Macmillan.
- Rost, J., & von Davier, M. (1994). A conditional item fit index for Rasch models. *Applied Psychological Measurement*, 18, 171–182.
- Sammons, P. Hillman, J., & Mortimore, P. (1995). *Key characteristics of effective schools: A review of school effectiveness research*. London: Office for Standards in Education.
- Scantlebury, K., Boone, W., Kahle, J. B., & Fraser, B. J. (2001). Design, validation, and use of an evaluation instrument for monitoring systemic reform. *Journal of Research in Science Teaching*, 38, 646–662.
- Scheerens, J. (1992). *Effective schooling: Research, theory and practice*. London: Cassell.
- Scheerens, J., & Bosker, R. (1997). *The foundations of educational effectiveness*. Oxford: Pergamon.
- Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modelling and confirmatory factor analysis results: A review. *The Journal of Educational Research*, 99, 323–338.
- Scientific Software International. (2011). *IRTPRO users' guide*. Lincolnwood, IL: Author.
- Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, 77, 454–499.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory, a primer*. London: Sage.
- Sijtsma, J. (1997). *Balans van het taalonderwijs aan het einde van de basisschool: Uitkomsten van de tweede taalpeiling einde basisonderwijs* [Achievements in mother tongue at the end of primary education]. Arnhem: Cito.
- Slavin, R. (1996). *Education for all*. Exton, PA: Swets & Zeitlinger.
- Smith, T. W., Baker, W. K., Hattie, J. A. C., & Bond, L. (2008). A validity study of the certification system of the National Board for Professional Teaching Standards. In L. Ingvarson & J. A. C. Hattie (Eds.), *Assessing teachers for professional certification: The first decade of the National Board for Professional Teaching Standards* (pp. 345–378). Oxford: Elsevier.
- Stieger, L. E. W. (2013). Beginning teachers: Issues of the impact of preparation and context on practice. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, 74, 2013.
- Teodorović, J. (2011). Classroom and school factors related to student achievement: What works for students? *School Effectiveness and School Improvement*, 22, 215–236.
- Tomlinson, C. (2003). *Fulfilling the promise of the differentiated classroom: Tools and strategies for responsive teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Tomlinson, C. A. (2005). Grading and differentiation: Paradox or good practice? *Theory into Practice*, 44, 262–269.
- Van de Grift, W. J. C. M. (2007). Quality of teaching in four European countries: A review of the literature and application of an assessment instrument. *Educational Research*, 49, 127–152.
- Van de Grift, W. J. C. M. (2013). Measuring teaching quality in several European countries. *School Effectiveness and School Improvement*. Advance online publication. doi:10.1080/09243453.2013.794845
- Van de Grift, W. J. C. M., & Helms-Lorenz, M. (2012). *Vaardigheid en ervaring van leraren* [Teaching skills and teacher experience]. Manuscript submitted for publication.
- Van de Grift, W. J. C. M., Helms-Lorenz, M., & Maulana, R. (2013). *Observeren van leraren in opleiding* [Observing student teachers]. Manuscript submitted for publication.

- Van de Grift, W. J. C. M., Van der Wal, M., & Torenbeek, M. (2011). Ontwikkeling in de pedagogisch didactische vaardigheid van leraren in het basisonderwijs [Development of teachers' teaching skills in primary education]. *Pedagogische Studiën*, 88, 416–432.
- von Davier, M. (1994). WINMIRA 2001 [Computer software]. Kiel, Germany: IPN.
- Wang, M. C., Reynolds, M. C., & Walberg, H. J. (1995). Serving students at the margins. *Educational Leadership*, 52(4), 12–17.
- Willms, J. D., & Somer, M.-A. (2001). Family, classroom, and school effects on children's educational outcomes in Latin America. *School Effectiveness and School Improvement*, 12, 409–445.
- Worthen, B. R., Sanders, J. R., & Fitzpatrick, J. L. (1997). *Program evaluation: Alternative approaches and practical guidelines*. White Plains, NY: Longman.
- Yair, G. (2000). Educational battlefields in America: The tug-of-war over students' engagement with instruction. *Sociology of Education*, 73, 247–269.

Appendix 1. Comparison of teaching behaviour domains between Danielson (2013), Pianta and Hamre (2009), and Van de Grift (2007).

Danielson (2013)	Pinta and Hamre (2009)	Van de Grift (2007)
Domain 2: Classroom Environment	Domain 1: Emotional supports	
2a. Creating an Environment of Respect and Rapport	Positive climate	<i>Domain 1: Safe and stimulating learning environment</i>
2b. Establishing a Culture for Learning	Negative climate	<i>Domain 1: Safe and stimulating learning environment</i>
	Sensitivity	
	Regard for student perspective	
2c. Managing Classroom Procedures	Domain 2: Classroom organization	<i>Domain 2: Efficient classroom management</i>
	Behavioural management	
2d. Managing Student Behaviour	Productivity	<i>Domain 2: Efficient classroom management</i>
2e. Organizing Physical Space		
Domain 3: Instruction	Domain 3: Instructional support	
3a Communicating With Students		<i>Domain 3: Clear instruction</i>
3b Using Questioning and Discussion Techniques	Concept development	<i>Domain 4: Activating learning</i>
	Quality of feedback	
3c Engaging Students in Learning	Language modelling	<i>Domain 4: Activating learning</i>
3d Using Assessment in Instruction		<i>Domain 3: Clear instruction</i>
3e Demonstrating Flexibility and Responsiveness		<i>Domain 4: Activating learning</i>
	Instructional learning formats (Domain 2)	<i>Domain 5: Differentiation</i> <i>Domain 6: Teaching learning strategies</i>