# Student perceptions of teaching quality in five countries: A partial credit model approach to assess measurement invariance

van der Lans, R.M.; Maulana, R.; Helms-Lorenz, M.; Fernández-García, C.-M.; Chun, S.; de Jager, T.; Irnidayanti, Y.; Inda-Caro, M.; Lee, O.; Coetzee, T.; Fadhilah, N.; Jeon, M.; Moorer, P.

**DOI**
10.1177/21582440211040121

**Publication date**
2021

**Document Version**
Final published version

**Published in**
SAGE Open

**License**
CC BY

*Original Research*

SAGE Open
July-September 2021: 1–20

DOI: 10.1177/21582440211040121
journals.sagepub.com/home/sgo
SAGE

# Student Perceptions of Teaching Quality in Five Countries: A Partial Credit Model Approach to Assess Measurement Invariance

**Rikkert M. van der Lans[1,2], Ridwan Maulana[1], Michelle Helms-Lorenz[1], Carmen-María Fernández-García[3], Seyeoung Chun[4], Thelma de Jager[5], Yulia Irnidayanti[6], Mercedes Inda-Caro[3], Okhwa Lee[7], Thys Coetzee[5], Nurul Fadhilah[8], Meae Jeon[4], and Peter Moorer[1]**

**Abstract**

This study examines measurement invariance of student perceptions of teaching quality collected in five countries: Indonesia (n students = 6,331), the Netherlands (n students = 6,738), South Africa (n students = 3,422), South Korea (n students = 6,997) and Spain (n students = 4,676). The administered questionnaire was the My Teacher Questionnaire (MTQ). Student perceived teachers' teaching quality was estimated using the partial credit model (PCM). Tests for differential item functioning (DIF) were used to assess measurement invariance. Furthermore, if DIF was found, it was explored whether an application of a quasi-international calibration, which estimates country-unique parameters for DIF items, can provide more valid estimates for between-country comparisons. Results indicate the absence of non-uniform DIF, but presence of uniform DIF among most items. This suggests that direct comparisons of raw mean or sum scores between countries is not advisable. Details of the set of invariant items are provided. Furthermore, results suggest that the quasi-international calibration is promising, but also that this approach needs further exploration in the context of student perceptions of teaching quality.

**Keywords**

teaching, education, social sciences, measurement and scaling methods, research methods, social sciences, educational measurement & assessment, education, social sciences, reliability and validity, research methods, social sciences, student perceptions, teaching quality

## Introduction

This study examines measurement invariance (MI) of student perceptions of teaching quality. National studies conducted in different countries support the validity to use student perceptions to describe and study variation in teaching quality (e.g., Downer et al., 2015; Ferguson, 2012; Maulana & Helms-Lorenz, 2016; Sauerwein & Theis, 2021; van der Lans et al., 2019; Wagner et al., 2013). However, these studies do not indicate how such descriptions and results compare between countries. The aim of this study is to explore MI of student perceptions in five different countries to reveal a potential indication of how results obtained with student perceptions gathered through surveys compare internationally.

To date, studies examining measurement invariance of student perceptions of teaching quality are relatively rare. Notable exceptions are the studies by André et al. (2020) and Scherer et al. (2016). These studies report evidence of

[1]University of Groningen, The Netherlands
[2]Leiden University Medical Center, The Netherlands
[3]University of Oviedo, Spain
[4]Chungnam National University, Daejeon, South Korea
[5]Tshwane University of Technology, Pretoria, South Africa
[6]State University of Jakarta, Indonesia
[7]Chungbuk National University, Cheongju, South Korea
[8]University of Indonesia, Jakarta, Indonesia

**Corresponding Authors:**
Rikkert M. van der Lans, LUMC-Curium, Endegeesterstraatweg 27, 2342 AK Oegstgeest, The Netherlands.
Email: r.m.van_der_lans@lumc.nl

Ridwan Maulana, Department of Teacher Education, University of Groningen, Grote Kruistraat 2/1, 9712 TS Groningen, The Netherlands.
Email: r.maulana@rug.nl

partial-invariance between countries. More specific, André et al. (2020) and Scherer et al. (2016) found evidence supporting (partial) metric invariance but no support for scalar invariance. Both studies applied the Multiple Group Confirmatory Factor Analysis (MGCFA) method which is rooted in the factor analysis framework. A novelty of this study is that it applies the Partial Credit Model (PCM; a polytomous Rasch model) to examine MI of student perceptions of teaching quality.

Masters's (1982) PCM, and Muraki's (1992) Generalized (G)PCM, are popular methods for the assessment of MI of cognitive tests in the International large-scale assessments (ILSA's), like the Program for International Student Assessment (PISA), Progress in International Reading and Literacy Study (PIRLS), and the Trends in Mathematics and Science Study (TIMSS). The popularity of (G)PCM in ILSA's might be explained by the flexibility it offers for international comparisons. Specifically, (G)PCM allow researchers to relate scores on one instrument to those of another, which techniques are referred to as "scaling to achieve comparability" or "linking" (Kolen & Brennan, 2014). In ILSA's between-country comparisons are challenged by variation in curricula. It is impossible to administer the exact same item content in all countries due to variation in curricula. Therefore, linking is used to enhance international comparisons (e.g., Oliveri & von Davier, 2011). Although linking is not unique to the (G)PCM, this model provides additional flexibility to applications of it (Kolen & Brennan, 2014). In the above traditional use, linking is used to increase comparability of cognitive tests of different content. Another potential benefit of linking are its applications to adjust for non-invariance of the same test or questionnaire administred in different countries (e.g., Oliveri & Von Davier, 2011, 2014). Given the high likelihood to find evidence of partial invariance in international comparisons of student perceived teaching quality, the second aim of this article is to further explore whether and how linking can benefit international comparisons of student perceived teaching quality in situations of non-invariance.

The research questions are as follows:

1. To what extent are scores of student perceptions of teaching quality invariant across countries?
2. How does perceived teaching quality in different countries compare?

## Background

### *Conceptualization of Teaching Quality*

This study applies a conceptualization of teaching quality that is grounded in the literature on teaching and teacher effectiveness (e.g., Hattie, 2008; Muijs et al., 2014; van de Grift, 2014). Studies on teaching and teacher effectiveness have repeatedly found some behaviors to be effective, meaning that they contribute to student learning and school success. Examples of such effective teaching behaviors include providing students with clear examples, having students think aloud, and requesting students to reflect on their learning approaches. In this study, manifestations of effective teaching behavior are conceptualized as representing indications of teaching quality.

The variety in effective teaching behaviors is typically categorized to and/or summarized by five to seven broader factors or domains (Bell et al., 2019; Muijs et al., 2014). Prior research in Indonesia, South Korea, the Netherlands, South Africa, Spain, and Turkey applied CFA and MGCFA and provides evidence that in all six countries the variety in effective teaching behaviors is well-represented by a six-factor structure (André et al., 2020; Inda-Caro et al., 2019; Maulana & Helms-Lorenz, 2016). These studies termed these factors as domains and the six domains are labeled as safe and stimulating learning climate, efficient classroom management, clear and structured instruction, activating teaching, teaching learning strategies, and differentiation. The domains and an example item related to each domain are presented in Table 1.

The present study extends on the work by André et al. (2020). More specifically, it introduces and examines the invariance of a complementary conceptualization. In this complementary conceptualization all effective teaching behaviors are hierarchically ordered along one latent continuum of teaching quality. This conceptualization is grounded in theories on teacher development proposed by Berliner (2004) and Fuller (1969). Theories on the development of teaching quality generally describe its acquisition as unfolding across one single continuum. Furthermore, the theories describe the continuum as a sequence of five phases (Berliner, 2004) or three stages (Fuller, 1969).

Van de Grift et al. (2011) used these theories on teacher development to logically derive a single continuum of effective teaching behaviors. Their proposed model matched the identified phases and stages described by studies on teacher development with the six domains of effective teaching. Based on this match, they hypothesized a hierarchical ordering of the six domains starting from those including the least complex teaching behaviors—the acquisition of which marks the novice teacher that starts learning to teach—and ending with most complex effective teaching behaviors—the acquisition of which marks the expert teacher. Being well aware of the natural deviations from such stage-like hierarchical orderings, Van de Grift et al. (2011) suggested that the ordering should be assessed by probability. Figure 1 sketches their proposed representation. In Figure 1, the *x*-axis represents the continuum of teaching quality and the *y*-axis represents the probability on manifestation of effective teaching behaviors in classrooms. The icon under the *x*-axis represents one specific teacher's location on the continuum. The probability to manifest effective teaching behavior increases if teaching quality increases. This is visualized by s-curved

**Table 1.** The Six Domains, Their Conceptualization, and One Example Item of the "My Teacher" Questionnaire.

| Domain | Conceptualization | Example of item |
|---|---|---|
| Safe learning climate | Includes indicators that the teacher can maintain a classroom environment characterized by respect and care. | My teacher ensures that I feel relaxed in class |
| Efficient classroom management | Includes indicators that the teacher clarifies working procedures, shape routines, and maintain rules to structure the students' learning environment | My teacher applies clear rules |
| Clear and structured instruction | Includes indicators that the teacher is able to clarify and structure explanations | My teacher uses clear examples |
| Activating teaching | Includes indicators that the teacher stimulate student thinking about the lesson content | My teacher motivates me to think |
| Teaching learning strategies | Includes indicators that the teacher stimulates meta-cognitive processing | My teacher teaches me to check my solutions |
| Differentiation | Includes indicators that the teacher is able to adapt explanations and assignments to individual student needs. | My teacher knows what I find difficult. |

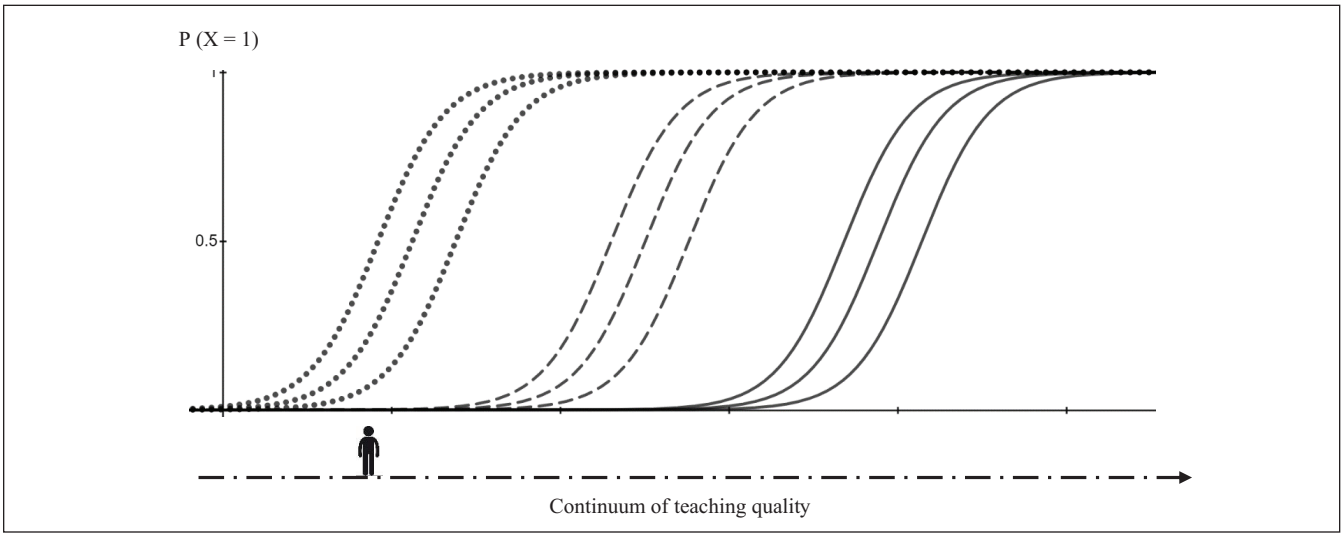


**Figure 1.** A non-empirical example of the theorized continuum of teaching quality.
*Note.* The s-curved dotted, dashed, and solid lines correspond to nine effective teaching behaviors associated with three domains of teaching quality. The icon positions one teacher on the continuum presumably one who is more at the beginning phases of teaching.

lines which indicate the probability that teachers manifest these effective teaching behaviors if a teachers location on the contiuum of teaching quality is known. The solid, dashed and dotted lines correspond to three domains. Let's assume the three domains are, from left to right, efficient classroom management, intensive and activating instructions, and differentiation. Then, the figure visualizes that an increase in the probability on manifestation of effective teaching behaviors in the domain intensive and activating instruction is predicted to be conditional on the probability on manifestation of effective teaching behaviors in the domain efficient classroom management.

Evidence related to conceptualization has been gathered in multiple studies and using a mixture of classroom observation and student questionnaire methods. Evidence obtained with both methods confirmed and further specified this hierarchical ordering in effective teaching behaviors (Maulana et al., 2015a; van de Grift et al., 2014; van der Lans et al., 2015, 2017, 2018, 2019). The ordering in domains approximately follows that presented in Table 1, with the exception of the final two domains. The questionnaire method estimates the domain teaching learning strategies as most complex, whereas the observation method follows the ordering as presented in Table 1. The current evidence-base is, however, mostly restricted to the Dutch context only. Notable exceptions are Indonesia (Maulana et al., 2015b), Cyprus (e.g., Kyriakides et al., 2018), and Turkey (Telli et al., 2020). To date, no studies have addressed the international invariance of the ordering in effective teaching behaviors.

## *Student Perceptions of Teaching Quality*

This study examines teaching quality as perceived by students. The term "perception" highlights that students' item

scores reflect their subjective experiences in the corresponding teachers' classes. This means that any two students in the same class can have different experiences and, thus, different perceptions. When this study mentions about the probability that teachers display effective teaching behaviors, this probability is estimated based on students' perceptions. When the study refers to estimations of teaching quality, it, in all instances, refers to the student perceived teaching quality.

Empirical evidence indicates that student perceptions vary primarily as a function of teachers' teaching quality (e.g., van der Lans & Maulana, 2018; van der Scheer et al., 2019; Wagner et al., 2013). Concerns with student perceptions mostly involve the potential for bias (e.g., Marsh & Roche, 2000; Spooren et al., 2013). Unlike classroom observers, students are not trained to score teaching quality using the predetermined standards. It is unclear which norms or standards students apply when scoring behaviors of their teachers. As will be discussed somewhat later in the article, the present examination of MI may provide some insights about whether the strength of student perception biases varies between countries.

## *Partial Credit Model: A Polytomous Rasch Model Approach to Study Cross-Country Comparisons*

Our prior research applied the Rasch model to gather evidence supporting an ordering in effective teaching behaviors. Mathematically, the Rasch model can be expressed as (Rasch, 1960):

$$P\left(X_{pi}=1\right)=\frac{e^{\left(\beta_p-\delta_i\right)}}{1+e^{\left(\beta_p-\delta_i\right)}} \quad (1)$$

Where $\beta_p$ estimates the student perceived position of the teacher on the continuum of teaching quality and $\delta_i$ estimates the location of effective teaching behaviors on the same continuum. Furthermore, the $\delta_i$ expresses what increase in teaching quality is predicted if teachers successfully display the effective teaching behavior *i*. Note that "successfully" is here defined by the students' subjective impression of the teacher's behavior and not to some objective norm.

The Rasch model is applicable to dichotomous item responses. The PCM extends on the Rasch model by introducing an item step parameter $\delta_{ik}$ (Masters, 1982). The PCM conceptualizes the Likert-type scale as consisting of *m* categories and of *m*−1 item steps. Item steps reflect the process of "stepping" from the lower to the next one-point higher item response category, such that $\delta_{ik}$ predicts what increase in teaching quality is associated with a one-point increase on the Likert-type scale response (i.e., step *k*) on item *i*. Mathematically, the PCM can be expressed as:

$$P\left(X_{pi}=1\right)=\frac{e^{\left(\beta_p-\delta_{ik}\right)}}{1+\left(\beta_p-\delta_{ik}\right)} \quad (2)$$

By using the PCM, the study keeps connection with multiple prior within-country studies indicating that students' perceptions of effective teaching behavior fit the Rasch model (Bacci & Caviezel, 2011; Bradley et al., 2006; Kyriakides et al., 2009; Maulana et al., 2015a; van der Lans et al., 2015). Also, it can provide a complementary perspective with prior studies that used (MG)CGA (e.g., André et al., 2020; Scherer et al., 2016).

*Rasch-type models and factor analytic models.* Several popular software packages like Mplus (Muthén & Muthén, 2019) and mirt (Chalmers, 2012) enable researchers to rescale parameters estimated using confirmatory factor analysis (CFA) into PCM parameters. These possibilities may give the impression that the two models themselves are identical. However, despite being mathematically identical, factor analytic models and Rasch-type models are conceptually different. The difference becomes most tangible in how the two models estimate model-data fit. Because factor analytic techniques enjoy considerable popularity, the above-mentioned conceptual difference is briefly explained.

Factor analytic fit tests are conceptually associated with classical test theory (CTT). Central to CTT is the argument that single observations are unreliable and that reliable estimates can be derived by averaging over multiple parallel observations (Graham, 2006). Factor analysis treats items as potentially parallel observations associated to one (or more) common factor(s). Expressed in a variance–covariance matrix, factor analytic fit tests assess the prediction of uniform item-covariance. Misfit to a one-factor model indicates that some item(s) are not essential tau-equivalent parallel. For more details, see Graham (2006) or Jöreskog (1971). Because factor analysis considers items to be parallel "replications" of the same latent factor/continuum, fit is typically estimated for the latent factor/continuum and variation in item parameters is typically interpreted as nuisance.

Contrary to factor analysis, Rasch-type models suggest that items vary in complexity ($\delta_i$) (more commonly referred to as difficulty) due to which items are no parallel observations (Brennan, 2010; Guttman, 1954). Expressed in a variance–covariance matrix, Rasch-type models fit tests assess the prediction that item-covariance decreases as a function of the distance between item (step) locations on the continuum (Browne, 1992; Guttman, 1954). This decreasing pattern is known as simplex structure and violates typical criteria set by factor analysis to assess parallelism (for details see: Jöreskog, 1978). In Rasch-type models, item parameters are no nuisance parameters which explains why fit is estimated per item and model fit is typically expressed by the joint item fit.

## *Partial Credit Model and Measurement Invariance: Interpretation and Meaning of Non-Invariance*

Rasch models typically examines MI in terms of Differential Item Functioning (DIF; French et al., 2019; Mazor et al.,

1994). In this study, two types of DIF are distinguished: uniform-DIF (U-DIF) and non-uniform-DIF (NU-DIF; Walker, 2011). U-DIF estimates between-country differences in the location of the same effective teaching behavior on the continuum. It signals that the teaching behavior (item) is associated with higher teaching quality in one country compared to another and that this difference is uniform across the continuum of teaching quality. NU-DIF, instead, estimates between-country differences in the slope or steepness with which the probability on an item response increases. It signals that the strenght of association of the teaching behavior (item) with the continuum of teaching quality varies between countries (Smith Walker, 2011).

Figure 2 visualizes two possible scenarios of NU-DIF which have different implications. When NU-DIF is constant (Scenario 3 in Figure 2), item slopes are parallel within countries but the strength of association between student perceived teaching behavior and the continuum of teaching quality varies between countries. When NU-DIF is inconstant (bottom scenario in Figure 2), then the item slopes within one or more of the countries are not parallel. This implies that in one or more countries no hierarchical ordering in teaching behaviors, as described above, can be derived.

Likewise, two scenarios can be derived for U-DIF. When U-DIF is constant (top scenario Figure 2), the students perceive all (or most) effective teaching behaviors as more complex. Because the direction and size of the shift in complexity is constant, we deem it more likely that this constant shift is due to differences between students' perceptions (e.g., between-country differences the subjective standards and norms applied by the students [i.e., strictness]), than that it represents differences in actual manifestations of effective teaching behaviors. Finally, when U-DIF is inconstant the evidence indicates between-country differences in how students hierarchically order the effective teaching behaviors. This scenario is likely when the actual manifestation of effective teaching behaviors in classrooms varies between countries.

We deem U-DIF as most plausible, but also argue that it has less severe consequences for measurement of teaching quality. Evidence suggesting between-country variation in the actual manifestation of teaching behaviors in classroom, for example, does not suggest real departures from the hypothesized continuum. It seems valid to apply linking in an attempt to improve between-country comparisons. The presence of NU-DIF, however, may result in more severe consequences. The slope-parameter provides information about an item's association with the latent continuum (Embretson & Reise, 2000; Fox, 2010) where lower slope parameters indicate lower association of an item with the continuum. Extending this interpretation, NU-DIF indicates that the student perceptions of effective teaching behaviors (items) are not related to the continuum (trait) in the same way across countries (Smith, 2002; Walker, 2011). Such differences in association seem unrelated to differences in actual teaching behaviors manifested in classrooms and introduces room to speculate about between-country differences in the impact of perception biases. Finally, NU-DIF may also indicate that the continuum derived by van de Grift et al. (2011), and which echoes prior theory on teacher development, does not generalize to other countries. The study will not apply linking to adjust (or correct) for NU-DIF.

## *Linking: Utility of Partial Credit Model Approach for International Empirical Research*

The PCM offers approaches to adjust for non-invariance in the form of linking (Ndosi et al., 2011; Oliveri & Von Davier, 2011, 2014; Tennant et al., 2004). Application of linking have been referred to as "quasi-international calibration" (Oliveri & Von Davier, 2011, 2014), "top-down purification" (Tennant et al., 2004), and "splitting of non-invariant items" (Ndosi et al., 2011). These differences in terminology express that the techniques are used for different reasons as well as that they differ in some technical details, nonetheless they follow the same underlying logic. In this study, quasi-international calibration is applied. Quasi-international calibration fixes invariant items and splits the non-invariant items by country (Oliveri & Von Davier, 2011, 2014). The resulting continuum combines emic effective teaching behaviors, which have culturally-general location in the hierarchy, and etic effective teaching behaviors, which have culturally-specific locations (Ndosi et al., 2011).

## *Context of the Current Study*

*The Netherlands.* International comparisons in secondary and primary education show that students attending Dutch schools perform above average, comparable to other high performing European and Asian educational systems (Mullis et al., 2016; Organisation for Economic Co-operation and Development [OECD], 2018). Teacher education for secondary education is divided into two different tracks. Teaching the lower levels of secondary education requires a second-degree teacher qualification, which takes four years of training (bachelor degree). Teaching the higher levels of secondary education requires a first degree teacher qualification; a subject-relevant master degree and an additional master at one of the university-based teacher education institutes. The first degree certification also allows teachers to teach the upper grades in higher levels of secondary education, i.e., higher vocational ("havo") and pre-university ("vwo"). The teaching profession does not have an above average status, and the quality of teachers is generally high with the large majority mastering the basic teaching skills well (OECD, 2016c).

*South Korea.* The South Korean educational system is among the top performing systems compared to most other countries in PISA and TIMSS (Mullis et al., 2016; OECD, 2018).

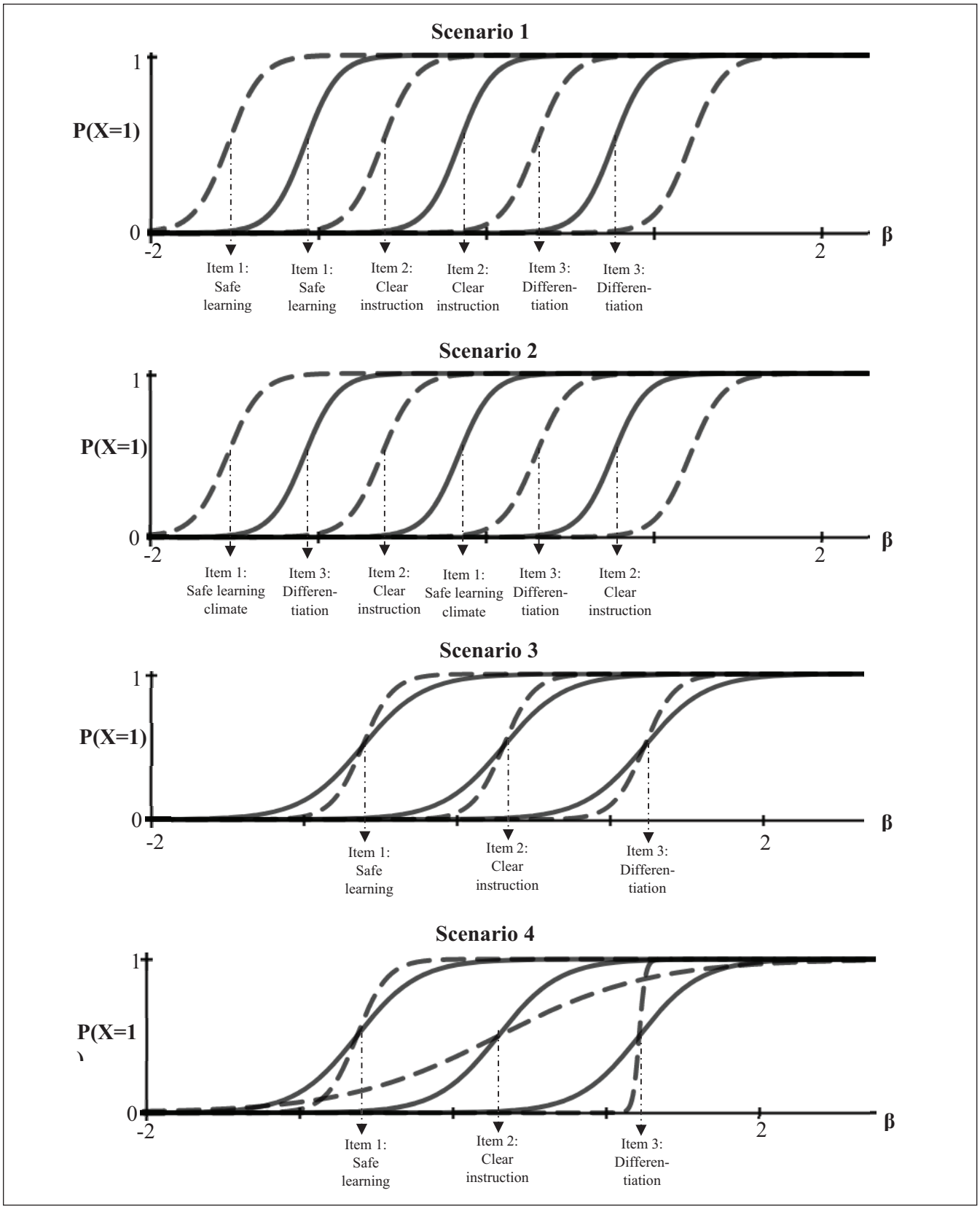


**Figure 2.** Four possible DIF-scenarios: (a) U-DIF with constant difference in location, (b) U-DIF with inconstant difference in location, (c) NU-DIF with constant difference in slope, and (d) NU-DIF with inconstant difference in slope.

*Note.* The dashed item characteristic curves refer to country A and the solid to country B. DIF = differential item functioning; U-DIF = uniform differential item functioning; NU-DIF = non-uniform differential item functioning.

Secondary school teacher training is offered as a four-year bachelor program which confers the second class certificate, later promoted to the first class by on-the-job experience, qualified to teach both at middle (7–9 grades) and high (10–12 grades) schools. For teaching at schools, the certificate holders should pass the highly competitive recruitment examination, a recent average of 10:1 pass rate, but securing a tenure job until 62 years (Korean education statistic center [KEDI], 2020). South Korea's student performance reveals a low percentage of underachieving students, and high percentages of excellent students. The South Korean system emphasizes on teaching quality and ongoing development in the teaching profession. Teaching profession is regarded as a highly-respected and high-status profession. Teachers are recruited from the top graduates, with strong financial and social incentives including social recognition as well as opportunities for career advancement and beneficial occupational conditions (Kang & Hong, 2008; OECD, 2016b).

*Indonesia.* The Indonesian educational system is among the lower performing countries in PISA (OECD, 2016a). Among many other components in the education system, Indonesian teachers play an important role in ensuring the success of the education system (Jalal et al., 2009). Teacher education for secondary education is offered as a four-year program at universities (Bachelor degree). Teacher certification is tied directly to their ability to demonstrate useful competencies, including meeting minimum levels of subject matter proficiency (de Ree, 2016b). Fasih et al. (2018), however, found that teacher certification is uncorrelated with student's learning outcomes. They suggest that this is due to the teacher training program which doesn't require implementation or demonstration of knowledge and skills *in* the classroom. Alternatively, de Ree (2016a) concludes that Indonesian teachers, though having completed a four-year bachelor degree program, have modest subject knowledge. Grounded on a country's ideal principle putting emphasis on respect for elderly and authority (Maulana et al., 2011), the teaching profession is regarded as a highly respected profession, but is not considered as having a high status. Therefore, improving the quality of education in Indonesia requires a broad agreement on the need to improve education quality and full commitment from all stakeholders, politicians, policymakers, unions, teachers, and parents.

*South Africa.* The South African educational system is developing, but currently its performance is from an international perspective ineffective. Based on TIMSS 2015, the country was ranked second last in mathematics and last science (Mullis et al., 2016). Moreover, of 139 participating countries, South Africa scored number 137 for overall quality of education (Baller et al., 2016). Teacher training programs consist of a four-year Bachelor degree course offered at higher education institutions. In addition, students qualified with specific content Bachelor degrees, for example, Engineers and Scientists, can complete a Post Graduate Diploma to become a qualified secondary school teacher. This Post Graduate Diploma equips potential teachers with competencies and pedagogical knowledge to teach diverse groups of students (Machingambi, 2020). Although significant improvements in basic and tertiary education is detected, the quality of education and teacher education is still not on par with other developing countries (van der Berg, 2015). For example, Taylor et al. (2013) showed that in six South African universities, only 6% of the curriculum for teacher training and development include how a teacher should teach a student to read. The education system still encounters various challenges which have been argued as related to the English second language instruction barrier, insufficient subject knowledge of some teachers, lack of accountability of teachers, frequent absenteeism of teachers from classes, and socioeconomic status of most students (Howie et al., 2012; Mbiti, 2016).

*Spain.* Spain performs around the average on PISA and TIMMS, but regional differences are relatively large (Hippe et al., 2018). The Southern region scores just above 470 points on PISA, whereas the capital of Madrid and the North-West score above 500 and closer to the Dutch average performance. Teacher training for primary education takes four years and is completed with a university degree (*Grado en Maestro de Educación Infantil o Primaria).* Teacher training for secondary education requires a relevant university degree (*Grado*) and an additional master in Teacher Training (Master's Degree in Teacher Training in Secondary and Upper Secondary Education and Vocational Training; Eurydice, 2019). The teaching profession has a reasonably high level of social prestige (over 70% of perceived social prestige scale). This image seems to be representative of the entire Spanish population, although research shows that the teachers might overestimate their reports (Centro de Investigaciones Sociológico [CIS], 2013; Fundación Europea Sociedad y Eduación, 2013; Gesellschaft für Konsum-, Markt- und Absatzforschung [GfK], 2018).

## Method

### *Sample and Data Management*

In total, five participating counties including Indonesia, South Korea, The Netherlands, South Africa and Spain collected survey data using the My Teacher Questionnaire (MTQ) from students of 4,918 teachers. Most survey data came from the Netherlands ($n = 3{,}519$ teachers). Teachers were approached to participate as part of country specific research projects. The year of country enrollment varied and available data spanned between one to four school years. Country samples were gathered using non-random sampling strategies, but all countries attempted to sample students and teachers from different regions to increase sample representativeness. The

Dutch data covers all 12 provinces. The Indonesian data covers provinces in the regions of Java, Sulawesi, Sumatra, and Kalimantan (the four main islands). The majority of the Spanish data are from the provinces Asturias and Galicia located in the North-West of Spain, plus a few teachers sampled in Andalusia (South of Spain). The South African data span the provinces of Gauteng, Kwazulu-Natal, and Mpumalanga. Finally, the South Korean data include students from the provinces Chungnam and Chungbuk.

*Inclusion and exclusion criteria.* In all countries, data of one school year were selected. Furthermore, a number of Dutch teachers ($n = 300$) proportionate to the number of teachers in other countries were randomly selected. Included school subjects were within the domains of languages, natural sciences, and social science and humanities. Subjects other than core subjects and which tend to be taught in alternative classroom settings, for example, physical education, music, and project-based education, were sampled but excluded from analyses. This selection leads to the final sample, which is referred to as the complete sample. The complete sample counted 1,456 teachers rated by 28,164 students from five different countries. Table 2 summarizes descriptive statistics of the country samples, including information on student gender, student age, subject taught, and class size.

Analyses were performed on two types of samples: (a) the complete sample and (b) the five randomly selected subsets. The complete sample has a nested design in which students grouped in the same class all score the same teacher. Analyses need to correct for the nested data structure (Hox, 2002). When corrections are not applied, the size of standard errors is likely underestimated which in turn increases the probability of type 1 errors. In the context of item fit tests, type 1 errors imply that we remove (or flag) items that actually fit. Hence, using the complete sample to assess item fit would imply an unnecessary strict assessment of item fit. Multilevel statistics can effectively remove bias due to the nested design, but these are not standard available in PCM estimation software. Therefore, the subsets were constructed by randomly selecting one student per class. These five subsets have equal sample size with $n$ equal to the number of teachers (Table 2). These subsets, also, effectively remove the nested design and provide more realistic estimations of item standard errors. Moreover, the selection of five random subsets provides the possibility to cross-validate findings. The complete sample is used to estimate the person parameters ($\beta_p$) and to describe, but not test, differences between countries.

*Missing values.* The overall number of missing values was low (0.8% of all item responses), but some of the returned questionnaires show multiple missing values. We excluded questionnaires showing more than five missing values (1.5% of all questionnaires), of which 11 questionnaires were from Indonesia, two from South Korea, 55 from the Netherlands, 341 from South Africa, and 29 from Spain. Reasons for why South Africa has the largest number of missing values in the questionnaires are unclear. Presumably, the reasons are likely related to the conditions of the students during the survey in the country which may include low literacy (difficulty in understanding certain questions), disruptions (surveys were done in the class of between 36 and 47 students), low familiarity with responding to surveys, limited resources (e.g., no pens or pencils), and the insufficient support from the teachers or administrators of surveys.

## Measurement Procedures and Model

*My Teacher Questionnaire (MTQ).* The MTQ was constructed to measure student perceptions of teaching quality. This questionnaire is based on previously validated versions (eg., Maulana et al., 2015a; van der Lans et al., 2015). This version of the MTQ comprises 41 items that operationalize six domains: safe learning climate, efficient classroom management, clear and structured instruction, activating teaching, teaching learning strategies, and differentiation (see also Table 1 in the background section). Response categories were provided on a 4-point Likert-type scale, ranging from 1 (never) to 4 (often), which were recoded into: 1 = 0, 2 = 1, 3 = 2, 4 = 3. Recoding was required for the intended PCM analysis.

*Translation procedure.* In the five countries, the questionnaire was translated from English to the target language and back-translated in accordance with the guidelines of the International Test Commission (Hambleton, 2001; van de Vijver & Tanzer, 2004). This procedure was recommended because it takes into account both the linguistic as well as the cultural and psychological aspects involved. The target language is as follows: Dutch for the Netherlands, Korean for South Korea, Bahasa Indonesia for Indonesia, English for South Africa, and Spanish for Spain. In each country, the translation and back-translation process involved two researchers highly knowledgeable about the technical and conceptual details of the MTQ and two university experts who are proficient in both English and the target languages. During the process, issues and discrepancies were discussed thoroughly and resolved subsequently by the core research team. Although the process was quite long and laborious, the issues discussed were relatively minor and revolved around choosing the most representative word equivalence and the accuracy of word choice. The research team confirmed the relevance of the MTQ items in their own national contexts, providing evidence for face validity.

*Measurement model.* This study applies the Partial Credit Model (PCM; Masters, 1982). The PCM is chosen because it (a) keeps connection with multiple prior within-country studies indicating that students' perceptions of effective teaching behavior fit the Rasch-type models (Bacci &

**Table 2.** Sample Descriptives for Each of the Five Countries.

| | Indonesia | | South Korea | | The Netherlands | | South Africa | | Spain | |
|---|---|---|---|---|---|---|---|---|---|---|
| Parameter | n | N | n | N | n | N | n | N | n | N |
| $N_{(ques)}$ | 6,329 | — | 6,983 | — | 6,672 | — | 4,107 | — | 4,650 | — |
| $N_{(teachers)}$ | 299 | 122,8015 | 336 | 247,837 | 300 | — | 270 | — | 251 | 278,414 |
| $N_{(questionnaires\ per\ teacher)}$ | 21.2 | | 20.8 | — | 22.5 | — | 16.5 | — | 18.6 | — |
| %female students | 60.1% | 50% | 57.9% | — | 52.6% | — | 60.8% | — | 50.6% | 48% |
| Mean age (years (*SD*) | 16.53 (1.00) | — | 16.39 (1.52) | | 13.85 (1.41) | | 15.27 (1.35) | — | 15.94 (1.53) | — |
| %public schools | — | 50.6% | 61.8 | 69.1% | 100.0 | — | 99.5 | — | 61.7 | 63.7% |
| % language subjects | 21.5 | — | 46.4 | — | 44.7 | — | 23.6 | — | 41.8 | 38.1% |
| % natural science subjects | 48.7 | — | 36.0 | — | 29.7 | — | 39.9 | — | 30.6 | 33.3% |
| Mean total score (*SD*) | 69.03 (13.28) | — | 82.65 (16.77) | — | 73.84 (18.13) | — | 74.77 (19.21) | — | 74.64 (13.87) | — |

*Note.* The n is sample and N is the countries total population. For Indonesia: Only certain background variables are available (http://publikasi.data.kemdikbud.go.id/uploadDir/isi_FBB7E3E1-3F01-49E6-B1BC-E1DA8E608D33.pdf). South Korea: Only certain national demographic information is available (https://kess.kedi.re.kr/index). The Netherlands: National demographic information listed in Table 2 is not publicly available. South Africa: National demographic information listed in Table 2 is not publicly available. Spain: % language subject = considering only compulsory subjects in Lower Secondary Education: 33.3%; in Upper Secondary Education; 42.85%. % natural science subject = considering only compulsory subjects in Lower Secondary Education; in Upper Secondary Education there are no natural Science subjects included in the compulsory ones (these subjects are only for those students who choose Scientific Upper Secondary Education but not for those who choose Humanities and Social Sciences; or Arts Upper Secondary Education (https://www.educacionyfp.gob.es/dam/jcr:957c29bb-ebd1-4e5b-9417-3d163cc32def/cifrasweb.pdf). Designs and random sample subsets.

Caviezel, 2011; Bradley et al., 2006; Kyriakides et al., 2009; Maulana et al., 2015a; van der Lans et al., 2015), (b) can be generalized to include a discrimination parameter (Muraki, 1992), which is important to assess NU-DIF, and (c) can handle items with different numbers of response categories. The latter two advantages anticipate on flexibility possibly required in future research.

## Analysis Plan

*Step 1: Model and item fit.* As a first step, the presence of the hierarchical ordering was evaluated in the separate countries. Tests assessing dimensionality involved (a) principal component analysis (PCA), (b) simplex analysis (Browne, 1992; Guttman, 1954), and (c) Mokken's H-coefficient (results only in Supplementary File Chapter 1; van der Ark, 2007).

PCA is not specifically developed to assess hierarchical ordering, but instead is a general factor analytic approach. It was estimated using the R package *psych* (Revelle & Revelle, 2015). Polychoric correlations were inserted instead of the default Pearson product correlations as recommended by Timmerman et al. (2018). To decide what the minimum number of factors was that still adequately represented the data we applied: (a) Horn's (1965) parallel analysis (PA) method, which selects the number of components with Eigenvalues higher than the Eigenvalues generated in a Monte Carlo simulation of equal sample size with random item responses and (b) Cattell's (1966) elbow rule, which states to retain the number of factors on the left side of the "elbow" in the scree plot. Instead of PCA, simplex analysis is specifically developed to assess hierarchical ordering (Browne, 1992; Guttman, 1954). The estimation of a simplex model is, however, currently only available via the FORTRAN program CIRCUM developed by Browne (1992). Although CIRCUM can estimate the simplex model, it requires researchers to constrain to item parameters to have equal distances on the latent measurement scale. This constraint is unnecessary strict but cannot be removed. CIRCUM provides just two absolute fit indices: (a) the root mean square error of approximation (RMSEA) and (b) the chi-square log-likelihood ratio test. Fit of the simplex model was assessed using RMSEA, where RMSEA $< 0.08$ indicated fair fit and RMSEA $<0.05$ indicated good fit (Browne & Cudeck, 1993). Finally, two coefficients of internal consistency, namely the lowest possible Split-half reliability and McDonald's (1999) omega coefficient, were estimated using the R package *psych* (Revelle & Revelle, 2015). In all analyses of dimensionality and/or internal consistency, the complete sample was used (see sample section).

As a second step, item fit was estimated using the Mean square (MS) item-infit and outfit coefficients. The traditionally advised cutoff criterion for MS infit and oufit is 1.20 (Bond & Fox, 2007), but more recent simulation studies show the necessity to accommodate criteria to the number of items and sample size (Seol, 2016). The number of items included is 41, the *n* ranges 251 to 336. Seol (2016) suggest cutoff values around 1.18 for these numbers. Given that 1.18 is close to the regularly advised cutoff by Bond and Fox (2007), it was decided to apply this regular cutoff $>1.20$. It should be noted that subsequent DIF analyses in step 2 apply stricter item fit criteria. Any false-positive item fit results at step 1 likely are corrected at step 2. Item fit was examined five times in five different subsets of the data (see sample section and Supplementary File: Chapter 1).

*Step 2. Evidence of MI.* U-DIF and NU-DIF were assessed using the R package *lordif* (Choi et al., 2011). *Lordif* expresses DIF using the *p*-value ($\chi^2$ difference test) and using pseudo-$R^2$ effect size measures. The combination of *p*-values and effect size measures gives superior control over potential type-1 errors (Choi et al., 2011). In this study, the DIF-effect size refers to McFadden's pseudo $R^2$. Cut-off criteria for the *p*-value and $R^2$ were estimated using a Multiple-Chain-Monte-Carlo (MCMC) simulation study (Choi et al., 2011). Exact cut-off criteria are reported in the Supplementary File, Chapter 1. DIF was assessed five times using five subsets of the data (see sample section).

*Validating DIF results.* False-positive DIF results can occur in samples that have different distributions of background variables. Imagine that an item has DIF for gender and that gender is unequally distributed among the countries. To validate the results in step 2, DIF analyses were conducted using a selection of the complete sample that matched the five country samples on student gender and student age. The selection of these two variables was based on preliminary DIF-analyses using the R package *psychotree* (Zeileis et al., 2009). Another not-matching sample of equal size was randomly selected. DIF was assessed in the matched and not-matched datasets using *lofdif*. Results indicated no evidence that DIF results were affected by differences in the distribution of background variables, thus, supporting the findings obtained in step 1. The complete procedure is reported in Supplementary File, Chapter 2.

*Step 3: Linking though quasi-international calibration.* To answer the second research question differences in country-average student perceived teaching quality were explored between the standard international calibration approach, which assumes that all items are invariant, compared to a quasi-international calibration approach. Differences between calibration methods were expected because of prior results that indicate partial measurement invariance (e.g., André et al., 2020; Scherer et al., 2016). In case that calibration results differed, model fit estimates were compared to indicate what, from a purely data-driven approach, calibration method to prefer.

*Quasi-international calibration methods.* Two approaches of quasi-international calibration were applied, namely concurrent and separate. In the concurrent approach, all item

parameters were estimated in one step by fixing the invariant items to be equal and estimating country-unique item parameters for non-invariant items (Oliveri & von Davier, 2011, 2014). The analysis was performed using the R packages *eRm* and applying default settings (Mair & Hatzinger, 2007). The separate calibration approach took two steps. First, item step parameters were estimated for the separate countries using the PCM function of the package *eRm* and applying the default settings. The output was provided to the R package *plink* (Weeks, 2010). Plink re-calibrates item parameters of one (focal) country onto another's country continuum using transformation constants that are estimated based on the invariant items. *Plink* offers four distinct methods to estimate transformation constants: mean-mean, the mean-sigma, the Haebrema, and the Stocking–Lord transformation (Kolen & Brennan, 2014; Weeks, 2010). The mean-mean and mean-sigma methods are known as the moment methods, and the Haebrema and Stocking–Lord methods as the item characteristic curve methods. The few available simulation studies indicate the item characteristic curve methods provide more accurate item parameters (Hanson & Béguin, 2002; Kilmen & Demirtasli, 2012; Kolen & Brennan, 2014). In this study, the stocking–Lord transformation was applied for separate calibration. Because we applied separate calibration with more than two countries, the countries needed to be chained. The chain applied in this study is: Netherlands-South Korea, South-Korea-Indonesia, Indonesia-South Africa, and South Africa-Spain.

Applications of concurrent and/or separate quasi-international calibration are relatively novel. Also, various psychometric models can be used, though the results might have different interpretations. Available evidence concerning the concurrent calibration method indicate that it is quite robust. Arai and Mayekawa's (2011) simulation study, for example, examined the number of invariant items required to validly perform concurrent calibration. Their results indicated that concurrent calibration may be valid with few, perhaps even less than five, invariant items. In an empirical study by Chen et al. (2009), this finding is corroborated. Another simulation study by Liu et al. (2011) examined whether the invariant items need to cover the complete continuum. Their results signal that this might not be a requirement.

*Fit of calibrations.* No uniform standard currently exists to estimate the fit of quasi-international concurrent or separate calibration. Prior work applied other psychometric models than the here applied PCM (Ndosi et al., 2011; Oliveri & Von Davier, 2011, 2014; Tennant et al., 2004) and each report another estimate of model and/or item fit. This study reports country-mean item and person MS-outfit statistics. The outfit-statistic equals the Chi-square value divided by its degrees of freedom (df). Outfit values of 1.00 indicate complete model fit and the further values depart from 1.00 the lower the model fit is. The country-mean outfit statistics are supplemented with the Minimum and Maximum to give an impression of the distribution. Unfortunately, the R package *plink* does not provide any item, person or model fit estimates. Hence, currently information about item, person, and model fit cannot be provided for the quasi-international separate calibration.

## Results

### *Step 1: Screening of Model and Item Fit in the Separate Country Data*

Results of the PA method and the simplex analysis are presented in Table 3. Guttman's simplex analysis indicates adequate fit of the data to the predicted simplex correlation structure in each country (RMSEA $< 0.08$). When applying Horn's PA method, the number of extracted factors varies but in all countries is greater than one. This was expected because the conceptualization predicts (six) local clustering's on the continuum. Furthermore, the PA method is sensitive to large sample size. In this study sample sizes ranged from $n = 6{,}983$ to $n = 4{,}107$. Using Cattell's elbow rule, the PCA scree plots (see Supplementary File Chapter 1) suggests the presence of one dominant factor within each country except perhaps for Spain. For the Spanish data, the second component is larger than 3.00, which is relatively large when compared to the first component (12.40). Simplex analysis, instead, suggests good fit of the Spanish data to the continuum (RMSEA $< 0.05$). Because simplex analysis was designed to estimate fit of a hierarchical item response pattern, the results for the Spanish data are deemed adequate. Internal consistency, as estimated by McDonald's omega and the lowest split half reliability, is high (see Table 3).

Four items were found to misfit the continuum in multiple countries using the MS-infit and MS-outfit. These items were not considered in the analysis of MI (for details, see Supplementary File, Chapter 1).

### *Step 2: Evidence of MI*

Table 4 summarizes the results of the NU-DIF and U-DIF. The columns indicate the two criteria, namely McFadden's pseudo R-square and the Chi-square test, and indicate whether the item was flagged for U-DIF and/or NU-DIF. TRUE means that an item was flagged more than once in the five samples and according to both criteria. Results show that none of the items are (repeatedly) flagged for NU-DIF, but also that all but four items are repeatedly flagged for U-DIF. The four invariant items are: "*My teacher makes sure that I pay attention*," *My teacher uses clear examples*," "*My teacher applies clear rules*," "*My teacher pays attention to me*." The Supplementary File Chapter 1 provides details of the item DIF results.

Table 5 summarizes the pooled item location parameters of the six domains. The domains "efficient classroom management," "clear and structured instruction," "activating

**Table 3.** Summary of Results for the One-Dimensionality Analysis of the MTQ Student Perception Survey for Indonesia, South Korea, Netherlands, South Africa, and Spain.

| | | | Parallel analysis | | Guttman's simplex | | Internal consistency | |
|---|---|---|---|---|---|---|---|---|
| Country | *n* | C | Observed | Simulated | RMSEA (90% CI) | $\chi^2$ (*df*) | omega | Lowest split-half |
| Indonesia | 6,329 | 1 | 15.25 | 1.15 | 0.068 | 23438.10 | 0.94 | 0.91 |
| | | 2 | 2.31 | 1.14 | (0.067; 0.069) | (776) | | |
| | | 3 | 1.78 | 1.13 | | | | |
| South Korea | 6,983 | 1 | 26.59 | 1.15 | 0.058 | 18825.24 | 0.98 | 0.95 |
| | | 2 | 1.78 | 1.13 | (0.057; 0.058) | (776) | | |
| | | 3 | 1.11 | 1.12 | | | | |
| Netherlands | 6,672 | 1 | 19.72 | 1.15 | 0.060 | 19676.53 | 0.96 | 0.93 |
| | | 2 | 2.24 | 1.13 | (0.059; 0.061) | (776) | | |
| | | 3 | 1.21 | 1.12 | | | | |
| South Africa | 4,107 | 1 | 19.76 | 1.20 | 0.055 | 11706.83 | 0.96 | 0.94 |
| | | 2 | 1.85 | 1.17 | (0.054; 0.056) | (776) | | |
| | | 3 | 1.67 | 1.16 | | | | |
| Spain | 4,650 | 1 | 12.40 | 1.18 | 0.045 | 7260.32 | 0.93 | 0.86 |
| | | 2 | 3.02 | 1.16 | (0.044; 0.046) | (776) | | |
| | | 3 | 1.30 | 1.15 | | | | |

*Note.* MTQ = My Teacher Questionnaire; RMSEA = root mean square approximation; CI = confidence interval.

teaching," "teaching learning strategies," and "differentiation" are similarly ordered along the continuum in all five countries. The domain "safe learning climate," however, clearly has different locations between countries. In the Netherlands and Spain (Europe), the domain "safe learning climate" is located at the start of the ordering and near "efficient classroom management." In South Africa, Indonesia, and South Korea, the domain is positioned third or fourth and located closer the domain "activating teaching". Furthermore, in South Korea and Indonesia (Asia), the specific items referring to "respect" are perceived by the students as located at the far end of the continuum of teaching quality. In terms of the conceptualization introduced above, this would imply that Indonesian and South Korean students associate these behaviors with "expert" teaching. This contrasts with the European students which position items referring to "respect" at the start of the continuum.

### *Step 3: Linking Through Quasi-International Calibration*

Table 6 summarizes differences in the country median (*Mdn*) and mean (*M*) of student perceived teaching quality using four different metrics: (a) raw sum scores, (b) standard international calibration (assuming all items to be invariant), (c) concurrent quasi-international calibration, and (d) separate quasi-international calibration with the Stocking–Lord transformation.

Pearson correlations indicate that the two quasi-international calibrations are similar, to the standard international calibration, $r_{(df = 28,724)} = 0.99$ and $r_{(df = 26,567)} = 0.87$ for the concurrent and separate calibration, respectively. Nonetheless, country average teaching quality estimates are different depending on the calibration methods. The quasi-international separate calibration has highest between-country discrimination (range 2.20 logits), followed by the quasi-international concurrent calibration (range = 1.84 logits) and the standard international calibration (range = 1.60 logits). The raw scores discriminate the least. When using raw teaching quality estimates, the lowest country average score falls within one (pooled) standard deviation of the highest country average score.

The concurrent quasi-international calibration has superior person fit estimates compared to the standard international calibration. The mean person MS-outfit ranges from 0.75 in South Korea to 1.47 in South Africa in the standard international calibration and from 0.93 in South Korea to 1.10 in South Africa in the concurrent calibration. Fit of the separate calibration method is unknown. Results of the separate quasi-international calibration were found to be sensitive to the ordering of the chain. If the chain is ordered differently, the results changes. Thus, the separate calibration may yield highest discrimination, but its results are unreliable. The method needs further development. Wright maps are presented in the Supplementary File at the end of chapters 3, 4, and 5. The Wright maps present a quick overview of the match between item locations and person locations on the continuum of teaching quality.

## Conclusions and Discussion

The current study aims to investigate measurement invariance (MI) of student perceptions of teaching quality across countries including Indonesia, South Korea, the Netherlands,

**Table 4.** Overview of Items Flagged for Uniform-DIF (U-DIF) and Non-Uniform DIF (NU-DIF). TRUE Means That Items Are Flagged in More Than One of the Five Subsets.

| Item: | My Teacher . . . | U-DIF $R^2$ | U-DIF $\chi^2$ | NU-DIF $R^2$ | NU-DIF $\chi^2$ |
|---|---|---|---|---|---|
| 1 | . . . helps me if I don't know. | TRUE | TRUE | FALSE | TRUE |
| 2 | . . . makes sure that others treat me with respect. | TRUE | TRUE | FALSE | TRUE |
| 3 | . . . makes sure that I use my time effectively. | TRUE | TRUE | FALSE | FALSE |
| 4 | . . . makes clear what I need to learn for a test. | TRUE | TRUE | FALSE | TRUE |
| 7 | . . . answers my questions. | TRUE | TRUE | FALSE | FALSE |
| 8 | . . . takes into account what I already know. | TRUE | TRUE | FALSE | FALSE |
| 9 | . . . makes sure that I treat others with respect. | TRUE | TRUE | FALSE | TRUE |
| 10 | . . . explains how I need to do things. | TRUE | TRUE | FALSE | FALSE |
| 11 | . . . makes sure that I know what to do. | TRUE | TRUE | FALSE | FALSE |
| 12 | . . . explains everything clearly to me. | FALSE | TRUE | FALSE | FALSE |
| 13 | . . . makes sure that I keep on working. | TRUE | TRUE | FALSE | FALSE |
| 14 | . . . explains the purpose of the lesson clearly. | TRUE | TRUE | FALSE | FALSE |
| 15 | . . . talks interestingly. | TRUE | TRUE | FALSE | FALSE |
| 17 | . . . teaches me to check my solutions. | TRUE | TRUE | FALSE | FALSE |
| 18 | . . . encourages me to think. | TRUE | TRUE | FALSE | FALSE |
| 19 | . . . makes clear to me why my answers are good or not. | TRUE | TRUE | FALSE | FALSE |
| 20 | . . . states clearly when assignments/tasks are due. | TRUE | TRUE | FALSE | FALSE |
| 21 | . . . prepares his/her lessons well. | TRUE | TRUE | FALSE | TRUE |
| 22 | . . . approaches me with respect. | TRUE | TRUE | TRUE | TRUE |
| 23 | . . . stimulates me to cooperate with my classmates. | TRUE | TRUE | FALSE | FALSE |
| 24 | . . . makes sure that I pay attention. | FALSE | FALSE | FALSE | FALSE |
| 25 | . . . uses clear examples. | FALSE | TRUE | FALSE | FALSE |
| 26 | . . . makes connections to what I already know. | TRUE | TRUE | FALSE | FALSE |
| 27 | . . . applies clear rules. | FALSE | FALSE | FALSE | FALSE |
| 29 | . . . tells how I should learn something. | TRUE | TRUE | FALSE | TRUE |
| 30 | . . . makes me feel self-confident with difficult tasks. | TRUE | TRUE | FALSE | FALSE |
| 31 | . . . motivates me to think. | TRUE | TRUE | FALSE | FALSE |
| 33 | . . . pays attention to me. | FALSE | FALSE | FALSE | FALSE |
| 34 | . . . states the lesson objectives. | TRUE | TRUE | FALSE | FALSE |
| 35 | . . . checks whether I have understood the content of the lesson. | TRUE | TRUE | FALSE | FALSE |
| 36 | . . . motivates me. | TRUE | TRUE | FALSE | FALSE |
| 37 | . . . knows what I have difficulty with. | TRUE | TRUE | FALSE | FALSE |
| 39 | . . . makes sure that I do my best. | TRUE | TRUE | FALSE | FALSE |
| 40 | . . . involves me in the lesson. | TRUE | TRUE | FALSE | TRUE |
| 41 | . . . helps me if I do not understand. | TRUE | TRUE | FALSE | FALSE |

*Note.* Item 12 was eventually not selected because it was flagged multiple times in one country for NU-DIF. This does not show in the table. DIF = differential item functioning; U-DIF = uniform differential item functioning; NU-DIF = non-uniform differential item functioning.

**Table 5.** Overview of DIF Between the Six Domains.

| Domain | Indonesia | South Korea | Netherlands | South Africa | Spain |
|---|---|---|---|---|---|
| Safe learning climate | 0.95 | 0.93 | 0.14 | 0.32 | 0.22 |
| Efficient classroom management | 0.80 | 0.62 | 0.18 | 0.25 | 0.39 |
| Clear and structured instructions | 0.68 | 0.78 | 0.42 | 0.17 | 0.43 |
| Activating instructions | 0.98 | 1.17 | 0.94 | 0.34 | 0.76 |
| Teaching learning strategies | 0.90 | 1.05 | 1.14 | 0.33 | 0.68 |
| Differentiation | 1.21 | 1.25 | 1.46 | 0.48 | 0.71 |

**Table 6.** Country Average Teaching Quality Scores and Fit of Teaching Quality Scores When Using the: (1) Raw Total Scores, (2) Standard International Calibration (Assuming Item Invariance), (3) the Concurrent Quasi-International Calibration, and (4) the Separate Quasi-International Calibration.

| Method by country | *n* | *Mdn* | *M* | *SD* | Mean item outfit | Mean person outfit | Min person outfit | Max person outfit |
|---|---|---|---|---|---|---|---|---|
| *Raw score* | | | | | | | | |
| Indonesia | 6,236 | 70 | 69.03 | 13.28 | — | — | — | — |
| South Korea | 6,717 | 82 | 82.65 | 16.77 | — | — | — | — |
| Netherlands | 6,614 | 76 | 73.84 | 18.13 | — | — | — | — |
| South Africa[a] | 2,817 | 76 | 74.77 | 19.21 | — | — | — | — |
| Spain | 4,196 | 75 | 74.64 | 13.87 | — | — | — | — |
| *International calibration* | | | | | | | | |
| Indonesia | 6,329 | 1.41 | 1.50 | 1.03 | 0.98 | 0.77 | 0.04 | 4.29 |
| South Korea | 6,983 | 2.50 | 3.10 | 2.07 | 0.98 | 0.75 | 0.04 | 4.65 |
| Netherlands | 6,672 | 1.93 | 1.97 | 1.49 | 0.98 | 1.06 | 0.04 | 4.41 |
| South Africa | 4,107 | 1.85 | 2.04 | 1.89 | 0.98 | 1.47 | 0.04 | 5.14 |
| Spain | 4,650 | 1.85 | 1.96 | 1.16 | 0.98 | 1.06 | 0.04 | 4.30 |
| *Quasi-international concurrent calibration* | | | | | | | | |
| Indonesia | 6,329 | 1.52 | 1.59 | 1.25 | 0.95 | 0.94 | 0.03 | 5.76 |
| South Korea | 6,983 | 2.92 | 3.43 | 2.40 | 0.95 | 0.93 | 0.05 | 7.63 |
| Netherlands | 6,672 | 2.02 | 2.08 | 1.50 | 1.00 | 1.00 | 0.21 | 7.32 |
| South Africa | 4,107 | 1.71 | 1.94 | 1.54 | 1.05 | 1.10 | 0.01 | 3.93 |
| Spain | 4,650 | 1.99 | 2.11 | 1.09 | 0.98 | 0.98 | 0.04 | 4.05 |
| *Quasi-international separate calibration* | | | | | | | | |
| Indonesia | 6,329 | 1.68 | 1.72 | 0.90 | — | — | — | — |
| South Korea | 6,983 | 2.32 | 2.62 | 1.54 | — | — | — | — |
| Netherlands | 6,672 | 1.96 | 2.02 | 1.49 | — | — | — | — |
| South Africa | 4,107 | 0.36 | 0.46 | 0.58 | — | — | — | — |
| Spain | 4,650 | 1.04 | 1.15 | 0.78 | — | — | — | — |

[a]Sample size for South Africa dropped substantially because of list-wise deletion. Please see the Supplementary File Chapter 2 for comments and thoughts on this.

South Africa and Spain. Furthermore, the study explores potential indication of differences in student perceived teaching quality across the five countries, based on results generated from the first aim.

## *Research Question 1*

The first research question is, "*To what extent are scores of student perceptions of teaching quality invariant across countries?* The results provide support for the hypothesized conceptualization in which all effective teaching behaviors are ordered along one latent continuum of teaching quality in the five countries. Of the four scenarios visualized in Figure 2, (one of) the top two may apply but the evidence does not suggest that the bottom two scenarios apply. Despite that item locations tend to vary between countries, five of the six domains are similarly ordered in all five countries. This suggests that in most instances, DIF has a 'local' effect (i.e., it mostly affects the location of effective teaching behavior within the domain). However, this result does not apply to the items related to the domain 'safe learning climate'. Items in this domain show considerable between-country variation in location on the continuum. Particularly, effective teaching behaviors within the domain safe learning climate are perceived as manifested by almost all teachers by Western-European students (Spain and the Netherlands), even those having comparatively poorly teaching quality. In Asian students' data (South Korea and Indonesia), these behaviors are perceived as manifested only by expert teachers. The findings, however, also indicate that in all countries students associate teaching behaviors within the domain "safe learning climate" with the continuum of teaching quality. Based on the interpretations for U-DIF provided in the background, we argue that the findings reflect between-country differences in the true manifestation by teachers either/or in the strictness with which students score these behaviors.

Although most items are flagged for U-DIF, we found four invariant items showing no NU-DIF and no U-DIF. This means that these items are statistically and content-wise interpreted similarly in the five countries. The four items are "*My teacher makes sure that I pay attention*," *My teacher uses clear examples*," "*My teacher applies clear rules*," and "*My teacher pays attention to me*." There is no

straightforward explanation for the invariance of these four statements. A simple observation is that all four items are relatively short. Using short sentences may decrease the potential errors during the translation process and the interpretation of meaning by translators and respondents. Short questions also reduce potential survey response fatigue, which can contribute to reducing response bias (Ben-Nun, 2011). Furthermore, two items use the word "attention," though in different contexts, and two items use the word "clear." Choosing right-on-target words for a questionnaire is essential to prevent ambiguity for diverse respondents (Belson, 1984). Finally, the items correspond to the first three domains: Safe learning climate, Efficient classroom management, and clear and structured instruction, meaning that the items appear to be concentrated on the less complex side of the continuum. Teaching behavior located at the less complex side of the continuum are predicted to be demonstrated by many teachers. Possibly, students are more acquainted with manifestations of these teaching behaviors and, therefore, could more accurately connect the item contents with their experiences.

## *Research Question 2*

The second research question is, "*How does perceived teaching quality in different countries compare?"* An answer to this question is not straightforward and depends substantially on the calibration method. If no psychometric models are applied, then student perceived teaching quality of South African teachers is the second highest and close to student perceived teaching quality of South Korean teachers. In the concurrent calibration, the student perceived teaching quality of South African teachers is the second lowest and not near the student perceived teaching quality of South Korean teachers. Looking at item-, person- and model-fit, the concurrent calibration seems to be the superior method compared to the standard international calibration. This finding is in line with prior research of Oliveri and Von Davier (2011, 2014). The finding of superior fit of concurrent calibration is achieved with only four anchor items, which echoes the findings by Arai and Mayekawa (2011) and Chen et al. (2009) indicating that concurrent calibration may be valid with few, perhaps less than five, anchor items. Also, the four anchor items were not representative of the complete continuum. The four items were located more or less in the lower end and center of the continuum. This finding is consistent with prior simulation studies suggesting that anchor items do not need to be representative for the complete continuum (Liu et al., 2011).

Overlooking the results of the quasi-international concurrent- and separate calibrations, then the ordering is relatively stable for the perceived teaching quality of South Korean, Dutch, and South African teachers. In all three calibration methods, South Korean teachers teaching quality is perceived highest by their students, and Dutch teachers teaching quality is perceived fairly high. South African students perceive the teaching quality in their lessons as relatively low. Although reasons for why students perceived their teachers more beneficially in South Korea and the Netherlands compared to South Africa are not identified in this study, discussing a conjecture about this may guide future research further.

The superior student perceived teaching quality of South Korean teachers seems to be consistent with the academic performance of their students as documented in ILSA's (OECD, 2018). The South Korean educational system is regarded among the top performing systems compared to most other countries in PISA and TIMSS (Mullis et al., 2016; OECD, 2018). South Korea's student performance reveals a low percentage of underachieving students, and high percentages of excellent students. The South Korean system emphasizes teaching quality and ongoing development in the teaching profession. Teachers are recruited from the top graduates, with strong financial and social incentives including social recognition as well as opportunities for career advancement and beneficial occupational conditions (Kang & Hong, 2008; OECD, 2016b). These personal and contextual factors pertaining to South Korean schools may likely contribute to their academic excellence, which could partly be reflected in this study by students' perception of the their teachers' teaching quality.

Similarly, the position of the Dutch teachers is consistent with the academic performance of their students as documented in ILSA's (OECD, 2018). In general, the quality of teachers is generally high with the large majority mastering the basic teaching skills well (OECD, 2016c). Teacher qualification in The Netherlands follows a relatively high level of academic loading. Teaching the higher levels of secondary education, i.e., higher vocational ("havo") and pre-university ("vwo"), requires a first degree teacher qualification (also known as academic teacher qualification). This qualification is obtained with a subject-relevant master degree in addition with a master at one of the university-based teacher education institutes. The second degree teacher qualification takes four years, but does not require a subject relevant master degree. For the Dutch sample, the years of teaching experience are known (this is unknown for all other countries). The number of beginning teachers included in the Dutch sample is relatively large and, thus, likely deviating from the other country samples. The Dutch teachers' age (likely somewhat younger) might be argued to have contributed to explaining a relatively high student perceived teaching quality, though most studies indicate that beginning teachers have lower teaching quality (Kini & Podolsky, 2016).

The comparatively poor performance of South African teachers is also consistent with results of student academic performance documented in ILSA's (Mullis et al., 2016). The country has been continuing to work toward educational excellence, although basic infrastructure and cultural factors like multiple official languages remain a big challenge.

Students are generally instructed in English as a second language (Howie et al., 2012). Teacher training institutes and professional development are still relatively weak. A recent review of teacher training programs of six South African universities suggested that only 6% of the curriculum for teacher training and development include how a teacher should teach a student to read Taylor et al. (2013).

Finally, results for Spanish and Indonesian teachers varied between the concurrent and separate calibration methods, with the Spanish teachers being close to the Dutch teachers according to the concurrent calibration, but scoring much lower in the separate calibration and Indonesia ranked lowest in the concurrent calibration and third (and average) in the separate calibration. In ILSA's Spain performs around the average on PISA and TIMMS and Indonesia performs poorly compared to other countries in PISA (OECD, 2016a). Hence, the results of the concurrent calibration demonstrate more overlap with the outcomes of ILSA's compared to the results of the separate calibration. Yet, this overlap might also be explained by similarity in applied calibration methods. Calibration and linking methods applied by ILSA's are conceptually more comparable to concurrent calibration than separate calibration.

In sum, there is a tendency that results based on the concurrent calibration in terms of perceived teaching quality seem to be more consistent with results of ILSA's in terms of student academic performance. This tendency provides an important insight because teaching quality has been shown to be the most significant factor for student learning and outcomes (Hattie, 2008). Although it is tempting to view this tendency as evidence of the validity of concurrent calibration, we suggest that it is currently too early to make such conclusions and that further research on the stability and consistency of separate and concurrent calibration methods is required.

## *Limitations and Directions for Future Research*

Although the present study has multiple strengths, it is also subject to some limitations. This study relies on convenience sampling. The Dutch sample disproportionally includes perceptions of the younger students (mean age 13 years) and likely included a sample of younger teachers. The data from South Korea, South Africa, and Spain cover only several regions of the country. Hence, we caution against generalizations of findings until replications with more representative samples are available.

It was not possible to apply the linking while taking into account the nested data structure due to the limited availability of technical software for estimating such models currently. Although the random selection of the five sample subsets takes into account the hierarchical structure of the data in its unique way, it remains unknown to what extent the results will differ when between-teacher variance is modeled statistically. Future research is advised to further explore possibilities to apply linking of international data on perceived teaching quality and taking into account the nested data structure, when the technical support will be available.

The quasi concurrent- and separate calibration methods provided distinct results. This inconsistency complicates practical applications of the quasi-international calibration method. It remains inconclusive which of the two calibration methods, either concurrent or separate calibration, should be preferred to increase fairness of cross-country comparisons. The concurrent calibration is conceptually less complex, but it applies strict assumptions about the invariant items, which are assumed to have identical item location parameters between countries (Oliveri & Von Davier, 2011, 2014). This strict assumption does not apply to the separate calibration method (Stocking & Lord, 1983). From an applied perspective, our findings indicate that the separate quasi-international calibration has largest impact on the country comparison, but its fit is unknown and the outcomes are dependent on the applied chain sequence. Hence, the present study suggests that both methods require further development before this approach can be applied to data about perceived teaching quality.

## *A Final Note*

The present study is part of a larger project that attempts to construct an infrastructure that can be used to measure effective teaching globally and to use this infrastructure to report results concerning country-average differences in teaching quality. The infrastructure includes countries of different cultural values, which obviously creates a need to maximize flexibility while keeping with important principles of measurement. Results show the complexity of building this type of infrastructure and at the same time underline its importance for the field of teaching and educational effectiveness in general. Currently, most empirical evidence is accumulated based on research using raw mean and sum scores of teaching quality. Our results suggest that these raw scores might be biased estimators of teaching quality. Furthermore, the study suggests that bias might, at least partially, be corrected by using a quasi-international calibration method. Whether the application of these methods will lead to novel or alternative insights about teaching and its effectiveness remains inconclusive. We will continue to build on this infrastructure to better understand teaching effectiveness and how to measure it globally.

### Author's Note

University of Groningen led the research project. Authors from partner universities contributed equally to this work.

### Acknowledgments

We thank all schools and students in the participating countries for their committed participation in this study. We thank Anna Meijer for her contribution to the data cleaning.

## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by the Netherlands Initiative for Education Research (NRO) through the Differentiation in teaching from the International Perspective (project number: 405-15-732), the Dutch Ministry of Education through the induction of Dutch beginning teachers project (project number: OCWOND/OD8-2013/45916 U), the Korean Research Fund through the Study for Improving Teaching Skill by Classroom Observation Analysis (project number: 2017S1A5A2A03067650), and the Directorate General of Higher Education of Indonesia (project number: 27/SP2H/DRPM/LPPM-UNJ/2018).

## Ethical Statement

The Institutional Review Board (IRB) of the Department of Teacher Education was established in January 2017. Research projects which were started before this official installation of the IRB did not require an approval from the IRB. All research projects before this date were reviewed and approved by the Director of the department. The current study was started at the end of 2014. Although an IRB did not exist yet during that time, studies conducted within the department followed the Netherlands Code of Conduct for Academic Practice (2014) and the Code of Ethics for research in the Social and Behavioral Sciences Involving Human Participants (2016).

## ORCID iDs

Rikkert M. van der Lans https://orcid.org/0000-0001-9108-645X

Yulia Irnidayanti https://orcid.org/0000-0001-9458-7532

## Supplemental Material

Supplemental material for this article is available online.

## References

André, S., Maulana, R., Helms-Lorenz, M., Telli, S., Chun, S., Fernández-García, C. M., & Jeon, M. (2020). Student perceptions in measuring teaching behavior across six countries: A multi-group confirmatory factor analysis approach to measurement invariance. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.00273

Arai, S., & Mayekawa, S. I. (2011). A comparison of equating methods and linking designs for developing an item pool under item response theory. *Behaviormetrika*, *38*(1), 1–16.

Bacci, S., & Caviezel, V. (2011). Multilevel IRT models for the university teaching evaluation. *Journal of Applied Statistics*, *38*(12), 2775–2791.

Baller, S., Dutta, S., & Lanvin, B. (2016). *Global information technology report 2016*. Ouranos.

Bell, C. A., Dobbelaer, M. J., Klette, K., & Visscher, A. (2019). Qualities of classroom observation systems. *School Effectiveness and School Improvement*, *30*(1), 3–29. https://doi.org/10.1080/09243453.2018.1539014

Belson, W. A. (1984). The design and understanding of survey questions. *Journal of the Royal Statistical Society. Series A*, *147*(1), Article 105. https://doi.org/10.2307/2981742

Ben-Nun, P. (2011). Respondent fatigue. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 1–2). SAGE.

Berliner, D. C. (2004). Describing the behavior and documenting the accomplishments of expert teachers. *Bulletin of Science, Technology & Society*, *24*(3), 200–212. https://doi.org/10.1177/0270467604265535

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Taylor & Francis.

Bradley, K., Sampson, S., & Royal, K. (2006). Applying the Rasch rating scale model to gain insights into students' conceptualisation of quality mathematics instruction. *Mathematics Education Research Journal*, *18*(2), 11–26.

Brennan, R. L. (2010). Generalizability theory and classical test theory. *Applied Measurement in Education*, *24*(1), 1–21. https://doi.org/10.1080/08957347.2011.532417

Browne, M. W. (1992). Circumplex models for correlation matrices. *Psychometrika*, *57*(4), 469–497.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). SAGE.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29.

Chen, W. H., Revicki, D. A., Lai, J. S., Cook, K. F., & Amtmann, D. (2009). Linking pain items from two studies onto a common scale using item response theory. *Journal of Pain and Symptom Management*, *38*(4), 615–628. https://doi.org/10.1016/j.jpainsymman.2008.11.016

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). Lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of Statistical Software*, *39*(8), 1–30.

Centro de Investigaciones Sociológicas (CIS). (2013). *Barómetro febrero de 2013*. http://www.cis.es/cis/opencms/ES/NoticiasNovedades/InfoCIS/2013/Documentacion_2978.html

de Ree, J. J. (2016a). *How much teachers know and how much it matters in class: Analyzing three rounds of subject-specific test score data of Indonesian students and teachers* (World Bank policy research working paper 7556). World Bank.

de Ree, J. J. (2016b). *Indonesia-teacher certification and beyond: An empirical evaluation of the teacher certification program and education quality improvements in Indonesia* (No. 104599, pp. 1–76). World Bank. https://doi.org/10.1596/1813-9450-7556

Downer, J. T., Stuhlman, M., Schweig, J., Martínez, J. F., & Ruzek, E. (2015). Measuring effective teacher-student interactions from a student perspective: A multi-level analysis. *The Journal of Early Adolescence*, *35*(5-6), 722–758. https://doi.org/10.1177/0272431614564059

Embretson, S. E., & Reise, S. P. (2000). *Item response theory*. Lawrence Erlbaum.

Eurydice. (2019). *Teachers and education staff*. https://eacea.ec.europa.eu/national-policies/eurydice/content/teachers-and-education-staff-78_en

Fasih, T., Afkar, R., & Tomlinson, H. (2018). *Learning for all* (No. 29379). World Bank.

Ferguson, R. F. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan*, *94*(3), 24–28. https://doi.org/10.1177/003172171209400306

Fox, J. P. (2010). *Bayesian item response modeling: Theory and applications*. Springer Science & Business Media.

French, B. F., Finch, W. H., & Immekus, J. C. (2019). Multilevel Generalized Mantel-Haenszel For Differential Item Functioning Detection. *Frontiers in Education*, *18*, Article 1847. https://doi.org/10.3389/feduc.2019.00047

Fuller, F. F. (1969). Concerns of teachers: A developmental conceptualization. *American Educational Research Journal*, *6*(2), 207–226.

Fundación Europea Sociedad y Eduación. (2013). *El prestigio de la profesión docente en España. Percepcion y Realidad* [The prestige of the teacher profession in Spain: Perception and reality]. Fundación Europea Sociedad y Educacion y Fundación Botín.

Gesellschaft für Konsum-, Markt- und Absatzforschung (GfK). (2018). *Trust in professions*. https://www.nim.org/sites/default/files/medien/135/dokumente/2018_-_trust_in_professions_-_englisch.pdf

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement*, *66*(6), 930–944. https://doi.org/10.1177/0013164406288165

Guttman, L. L. (1954). A new approach to factor analysis: The Radex. In P. F. Lazersfeld (Ed.), *Mathematical thinking in the social sciences* (pp. 258–348). The Free Press.

Hambleton, R. K. (2001). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment*, *17*(3), 164–172. https://doi.org/10.1027/1015-5759.17.3.164

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, *26*(1), 3–24. https://doi.org/10.1177/0146621602026001001

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

Hippe, R., Jakubowski, M., & Araújo, L. (2018). *Regional inequalities in PISA: The case of Italy and Spain* (EUR 28868). Publications Office of the European Union. https://doi/org/10.2760/495702

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179–185.

Howie, S. J., Van Staden, S., Tshele, M., Dowse, C., & Zimmerman, L. (2012). *PIRLS 2011: South African children's reading literacy achievement report*. Centre for Evaluation and Assessment (CEA), University of Pretoria.

Hox, J. (2002). *Multilevel analysis techniques and applications*. Lawrence Erlbaum.

Inda-Caro, M., Maulana, R., Fernández-García, C. M., Peña-Calvo, J. V., del Carmen Rodríguez-Menéndez, M., & Helms-Lorenz, M. (2019). Validating a model of effective teaching behaviour and student engagement: Perspectives from Spanish students. *Learning Environments Research*, *22*(2), 229–251. https://doi.org/10.1007/s10984-018-9275-z

Jalal, F., Muchlas, S., Chang, M. C., Stevenson, R., Ragatz, A. B., & Negara, S. D. (2009). *Teacher certification in Indonesia: A strategy for teacher quality improvement* (English). World Bank Group. http://documents.worldbank.org/curated/en/705901468283513711/Teacher-certification-in-Indonesia-a-strategy-for-teacher-quality-improvement

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*(2), 109–133.

Jöreskog, K. G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika*, *43*(4), 443–477.

Kang, N. H., & Hong, M. (2008). Achieving excellence in teacher workforce and equity in learning opportunities in South Korea. *Educational Researcher*, *37*(4), 200–207. https://doi.org/10.3102/0013189x08319571

Kilmen, S., & Demirtasli, N. (2012). Comparison of test equating methods based on item response theory according to the sample size and ability distribution. *Procedia-social and Behavioral Sciences*, *46*, 130–134. https://doi.org/:10.1016/j.sbspro.2012.05.081

Kini, T., & Podolsky, A. (2016). *Does teaching experience increase teacher effectiveness: A review of the research*. Learning Policy Institute.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices*. Springer.

Korean education statistic center (KEDI). (2020). https://kess.kedi.re.kr/

Kyriakides, L., Creemers, B. P., & Antoniou, P. (2009). Teacher behaviour and student outcomes: Suggestions for research on teacher training and professional development. *Teaching and Teacher Education*, *25*(1), 12–23. https://doi.org/10.1016/j.tate.2008.06.001

Kyriakides, L., Creemers, B. P., & Panayiotou, A. (2018). Using educational effectiveness research to promote quality of teaching: The contribution of the dynamic model. *ZDM*, *50*(3), 381–393. https://doi.org/10.1007/s11858-018-0919-3

Liu, J., Sinharay, S., Holland, P. W., Curley, E., & Feigenbaum, M. (2011). Test score equating using a mini-version anchor and a midi anchor: A case study using SAT® data. *Journal of Educational Measurement*, *48*(4), 361–379.

Machingambi, S. (2020). Academics' experiences of a post graduate diploma in higher education (PGDHE) programme: A case of one university in South Africa. *International Journal of African Higher Education*, *7*(1), Article 11553. https://doi.org/10.6017/ijahe.v7i1.11553

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, *20*(9), 1–20.

Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity, or innocent bystanders? *Journal of Educational Psychology*, *92*(1), 202–228. https://doi.org/10.1037/0022-0663.92.1.202

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Maulana, R., Helms-Lorenz, M., & van de Grift, W. J. C. M. (2015a). 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

Maulana, R., Helms-Lorenz, M., & van de Grift, W. J. C. M. (2015b). Pupils' perceptions of teaching behaviour: Evaluation of an instrument and importance for academic motivation in Indonesian secondary education. *International Journal of Educational Research*, *69*, 98–112. https://doi.org/10.1016/j.ijer.2014.11.002

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

Maulana, R., Opdenakker, M.-C., den Brok, P., & Bosker, R. (2011). Teacher-student interpersonal relationships in Indonesian secondary education: Profiles and importance to student motivation. *Asia Pacific Journal of Education*, *31*(1), 33–49. https://doi.org/10.1080/02188791.2011.544061

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1994). Identification of nonuniform differential item functioning using a variation of the Mantel-Haenszel procedure. *Educational and Psychological Measurement*, *54*(2), 284–291. https://doi.org/10.1177/0013164494054002003

Mbiti, I. M. (2016). The need for accountability in education in developing countries. *The Journal of Economic Perspectives*, *30*(3), 109–132. https://doi.org/10.1257/jep.30.3.109

McDonald, R. P. (1999). *Test theory: A unified treatment* (1st ed.). Lawrence Erlbaum.

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

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 international results in mathematics*. http://timssandpirls.bc.edu/timss2015/international-results/

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

Muthén, L. K., & Muthén, B. (2019). *"Mplus": The comprehensive modelling program for applied researchers—User's guide*.

Ndosi, M., Tennant, A., Bergsten, U., Kukkurainen, M. L., Machado, P., de la Torre-Aboki, J., . . . Hill, J. (2011). Cross-cultural validation of the Educational Needs Assessment Tool in RA in 7 European countries. *BMC Musculoskeletal Disorders*, *12*, Article 110. https://doi.org/10.1186/1471-2474-12-110

Organisation for Economic Co-operation and Development. (2016a). *Country note: Results from PISA 2015—Indonesia*. https://www.oecd.org/pisa/PISA-2015-Indonesia.pdf

Organisation for Economic Co-operation and Development. (2016b). *Education policy outlook: Korea*. http://www.oecd.org/education/Education-Policy-Outlook-Korea.pdf

Organisation for Economic Co-operation and Development. (2016c). *Netherlands 2016: Foundations for the future—Reviews of policies for national education*. https://doi.org/10.1787/9789264257658-en

Organisation for Economic Co-operation and Development. (2018). *PISA 2015: Results in focus*. http://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling*, *53*(3), 315–333.

Oliveri, M. E., & von Davier, M. (2014). Toward increasing fairness in score scale calibrations employed in international large-scale assessments. *International Journal of Testing*, *14*(1), 1–21. https://doi.org/10.1080/15305058.2013.825265

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.

Revelle, W., & Revelle, M. W. (2015). *Package "psych.."* The Comprehensive R Archive Network.

Sauerwein, M., & Theis, D. (2021). New ways of dealing with lacking measurement invariance. In A. Oude Groote Beverborg, T. Feldhoff, K. Maag Merki, & F. Radisch (Eds.), *Concept and design developments in school improvement research: Accountability and educational improvement* (pp. 63–82). Springer. https://doi.org/10.1007/978-3-030-69345-9_5

Scherer, R., Nilsen, T., & Jansen, M. (2016). Evaluating individual students' perceptions of instructional quality: An investigation of their factor structure, measurement invariance, and relations to educational outcomes. *Frontiers in Psychology*, *7*, Article 110. https://doi.org/10.3389/fpsyg.2016.00110

Seol, H. (2016). Using the bootstrap method to evaluate the critical range of misfit for polytomous Rasch fit statistics. *Psychological Reports*, *118*(3), 937–956. https://doi.org/10.1177/0033294116649434

Smith, L. L. (2002). On the usefulness of item bias analysis to personality psychology. *Personality and Social Psychology Bulletin*, *28*(6), 754–763. https://doi.org/10.1177/0146167202289005

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research*, *83*(4), 598–642. https://doi.org/10.3102/0034654313496870

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 201–210. https://doi.org/10.1177/014662168300700208

Taylor, N., van der Berg, S., & Mabogoane, T. (2013). *What makes schools effective: Report of the National School Effectiveness Study*. Pearson.

Telli, S., Maulana, R., & Helms-Lorenz, M. (2020). Students' perceptions of teaching behaviour in Turkish secondary education: a Mokken Scaling of My Teacher Questionnaire. *Learning Environments Research*, 1–23. https://doi.org/10.1007/s10984-020-09329-8

Tennant, A., Penta, M., Tesio, L., Grimby, G., Thonnard, J. L., Slade, A., . . . Tripolski, M. (2004). Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: The PRO-ESOR project. *Medical Care*, *42*(1), 137–148. https://doi.org/10.1097/01.mlr.0000103529.63132.77

Timmerman, M. E., Lorenzo-Seva, U., & Ceulemans, E. (2018). The number of factors problem. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (pp. 305–324). John Wiley & Sons. https://doi.org/10.1002/9781118489772

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

van de Grift, W., Helms-Lorenz, M., & Maulana, R. (2014). Teaching skills of student teachers: Calibration of an evaluation instrument and its value in predicting student academic engagement. *Studies in Educational Evaluation*, *43*, 150–159. https://doi.org/10.1016/j.stueduc.2014.09.003

van de Grift, W. J. C. M., Van der Wal, M., & Torenbeek, M. (2011). Ontwikkeling in de pedagogische didactische vaardigheid van leraren in het basisonderwijs [Primary teachers' development of pedagogical didactical skill]. *Pedagogische Studiën*, *88*, 416–432.

van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, *54*(2), 119–135. https://doi.org/10.1016/j.erap.2003.12.004

van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software*, *20*(11), 1–19.

van der Berg, S. (2015). What the Annual National Assessments can tell us about learning deficits over the education system and the school career. *South African Journal of Childhood Education*, *5*(2), 28–43.

van der Lans, R. M., & Maulana, R. (2018). The use of secondary school student ratings of their teacher's skillfulness for low-stake assessment and high-stake evaluation. *Studies in Educational Evaluation*, 58, 112–121. https://doi.org/10.1016/j.stueduc.2018.06.003

van der Lans, R. M., van de Grift, W. J. C. M., & van Veen, K. (2015). Developing a teacher evaluation instrument to provide formative feedback using student ratings of teaching acts. *Educational Measurement: Issues and Practice*, *34*(3), 18–27. https://doi.org/10.1111/emip.12078

van der 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. https://doi.org/10.1016/j.lindif.2017.07.007

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. *The Journal of Experimental Education*, *86*(2), 247–264. https://doi.org/10.1080/00220973.2016.1268086

van der Lans, R. M., van de Grift, W. J. C. M., & van Veen, K. (2019). Same, similar, or something completely different? Calibrating student surveys and classroom observations of teaching quality onto a common metric. *Educational Measurement: Issues and Practice*, *38*(3), 55–64. https://doi.org/10.1111/emip.12267

van der Scheer, E. A., Bijlsma, H. J., & Glas, C. A. (2019). Validity and reliability of student perceptions of teaching quality in primary education. *School Effectiveness and School Improvement*, *30*(1), 30–50. https://doi.org/10.1080/09243453.2018.1539015

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

Walker, C. M. (2011). What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. *Journal of Psychoeducational Assessment*, *29*(4), 364–376. https://doi.org/10.1177/0734282911406666

Weeks, J. P. (2010). plink: An R package for linking mixed-format tests using IRT-based methods. *Journal of Statistical Software*, *35*(12), 1–33.

Zeileis, A., Strobl, C., Wickelmaier, F., Komboz, B., & Kopf, J. (2009). *psychotree: Recursive partitioning based on psychometric models* (R package version 0.15-0). https://cran.r-project.org/web/packages/psychotree/index.html