Observed teaching behaviour in secondary education across six countries: measurement invariance and indication of cross-national variations

Ridwan Maulana, Stéfanie André, Michelle Helms-Lorenz, James Ko, Seyeoung Chun, Abid Shahzad, Yulia Irnidayanti, Okhwa Lee, Thelma de Jager, Thys Coetzee & Nurul Fadhilah


To link to this article: https://doi.org/10.1080/09243453.2020.1777170

© 2020 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

Published online: 05 Jun 2020.

Submit your article to this journal

Article views: 379

View related articles

View Crossmark data
Observed teaching behaviour in secondary education across six countries: measurement invariance and indication of cross-national variations

Ridwan Maulana, Stéfanie André, Michelle Helms-Lorenz, James Ko, Seyeoung Chun, Abid Shahzad, Yulia Irnidayanti, Okhwa Lee, Thelma de Jager, Thys Coetzee and Nurul Fadhilah

ABSTRACT
Effective teaching is a key factor in educational effectiveness and improvement. To facilitate the knowledge exchange regarding effective teaching in secondary education internationally, there has been a growing interest in using classroom observation measures. However, little is known regarding the comparability of observation measures across various national contexts (measurement invariance). Consequently, comparing effective teaching behaviour across contexts is compromised by issues of construct and measurement equivalence. In the present study, we investigated measurement invariance of the International Comparative Analysis of Learning and Teaching (ICALT) for measuring teaching behaviour across various national contexts including the Netherlands, South Korea, South Africa, Indonesia, Hong Kong-China, and Pakistan. Results showed that ICALT indicates full strict invariance in four out of the six countries, which provides a possibility to compare teaching behaviour in the Netherlands, South Korea, South Africa, and Indonesia. South Korea indicated the highest quality of teaching behaviour, while Indonesia indicated the lowest.

ARTICLE HISTORY
Received 29 April 2019
Accepted 29 May 2020

KEYWORDS
Classroom observation; measurement invariance; teaching behaviour

Introduction
International testing studies such as the Programme for International Student Assessment (PISA) of the Organisation for Economic Cooperation and Development (OECD) and the Trends in International Mathematics and Science Study (TIMSS) of the...
International Association for the Evaluation of Educational Achievement (IEA) have provided useful insights regarding trends in educational outcomes in secondary education around the world (Martin et al., 2016; Mullis et al., 2016; OECD, 2016). Based on these two large-scale studies, it is clear which countries are among the top performers in terms of student performance (i.e., above the OECD average) and which countries have more room to improve student performance (i.e., below the OECD average). These studies have been a major stimulus for many countries to do more research on educational effectiveness, especially in countries where the results are unexpectedly low (Reynolds et al., 2014).

However, differences in teaching performance of teachers across various countries are studied less frequently. This lack of knowledge might be caused by the fact that the current literature on teaching and teacher effectiveness is rather fragmented. Most of the studies were conducted in one-country, usually Western, settings (Klassen et al., 2018). Furthermore, teaching is typically measured using self-reports (i.e., student or teacher surveys), while using an observation instrument to measure teaching is rather scarce (Stroet et al., 2013). Self-reports are known to be more vulnerable for social desirability and response style bias (Maulana & Helms-Lorenz, 2016; van de Vijver & Tanzer, 2004). Although using observation instruments is practically costly and highly demanding, it is viewed as a more objective way of describing what happens in the classroom compared to surveys (Maulana & Helms-Lorenz, 2016).

Existing research from various cultural settings typically use different instruments to measure teaching practices. Ko and Sammons (2013) argued that “different instruments may measure different constructs and instruments can vary significantly in their external validity, that is, their applicability in different educational and national contexts” (p. 29). Consequently, comparing results from different contexts using various instruments is compromised by issues of construct and measurement equivalence. Using a similar instrument in different national contexts makes it possible to reduce some of these threats to the external validity. Despite the extensive work in the area of effective teaching, little is known about whether efforts to develop an internationally valid instrument to measure teaching behaviour across national borders are meaningful.

Differences in student performance as documented by the international testing studies demand explanations in terms of teaching behaviour. Research on teaching effectiveness indicates that about 15%–25% of differences in student achievement can be explained by the work of teachers (e.g., Aaronson et al., 2007; Bosker & Witziers, 1996; Houtveen & van de Grift, 2007a, 2007b; Houtveen et al., 2004, 2014; Rockoff, 2004; Roeleveld, 2003; Wijnstra et al., 2003). Hence, differences in the quality of teaching practices may explain differences in countries’ average educational outcomes. International comparative work has been highlighted as an important future research direction in educational effectiveness research (Reynolds et al., 2014). Regardless of the promising insights derived from international comparative research, knowledge transfer from one national context to another is challenged by the accuracy and sensitivity of the measure; whether or not the measure is well suited to the cultural contexts being compared (measurement invariance). Failing to prove measurement invariance in country comparison studies impedes the transferability of findings across contexts (Adamson, 2012).

Unravelling similarities and differences in teaching quality from a cross-national point of view is useful for the following reasons. First, it contributes to the advancement of
our knowledge of effective teaching behaviour. It can reveal similarities and differences in teaching quality between teachers across national contexts that otherwise would not be possible to identify from one-country studies (Adamson, 2012). Second, it offers ways to support international benchmarking efforts. Third, it provides hints for designing and learning from high-quality teacher improvement across various national contexts. Fourth, it provides information for schools to set appropriate criteria for (self-) evaluation. Additionally, it offers information for policy makers in the form of best practices across countries (Adamson, 2012).

Although research in primary education shows that teaching behaviour varies across several European countries (van de Grift, 2007, 2014), little is known about cross-national variations in teaching behaviour in secondary education. A first research of van de Grift et al. (2017) comparing teaching behaviour in the Netherlands and South Korea sheds light on the prospect of cross-national comparability in secondary education by showing that the used teaching behaviour measure showed sufficient cross-cultural invariance. However, their research employed the variable-centred approach assuming the response category measure to be continuous, while the measure itself has a categorical response structure. In educational research, the application of categorical multigroup confirmatory analyses is relatively new. Treating categorical variables as continuous may contribute to biased outcomes because the distance between categories is assumed to be equal, while in reality this is not the case, which is our first (methodological) contribution.

The current study offers insights into potential similarities and differences in teaching behaviour based on actual observations in authentic classroom settings. This is novel because most studies on measuring teaching quality typically rely on student perceptions or student performance (Skedsmo & Huber, 2019). Hence, our study delivers information regarding similarities and differences in teaching behaviour based on a more objective point of view. In addition, scientific studies measuring observed teaching quality involving multiple countries are currently not present, especially in the context of secondary education (cf. André et al., 2020; Fischer et al., 2019). This study is among the first to address this gap.

Markus (2016) noted that most of the world is not WEIRD (Western, Educated, Industrialized, Rich, and Democratic). This implies that variations in classroom practices are expected, especially when WEIRD and non-WEIRD contexts are included in the study. Because teaching is inherently cultural, in which values, beliefs, and goals play a role, similarities and differences in interpreting what constitutes effective teaching may differ depending on cultural contexts (Klassen et al., 2018). This suggests that some components of teaching are universally applicable across cultural contexts (generic), while other components are only applicable to certain cultural contexts (specific). The present study aims to extend and refine the previous work by examining whether the observation measure of teaching behaviour can be meaningfully compared (i.e., shows measurement invariance) in secondary education in six culturally contrasting countries: the Netherlands, South Korea, Indonesia, South Africa, Hong Kong-China, and Pakistan. To reach this aim, establishing measurement invariance is a pre-condition. Only with evidence of measurement invariance, comparing measures across countries can be deemed permissible. The aim of comparing measures across countries generates information about best practices, which opens opportunities to learn from each other. In
addition, the study intends to investigate potential differences in teaching behaviour across the six national contexts.

This study is guided by a methodological and a substantial research question:

(1) To what extent is the use of an observation instrument measuring effective teaching behaviour in secondary education invariant in six national samples (the Netherlands, South Korea, Indonesia, South Africa, Hong Kong-China, and Pakistan)?
(2) What are the differences and similarities regarding the six domains of effective teaching behaviour across the six national samples?

As mentioned earlier, the second research question can only be answered if there is sufficient evidence of measurement invariance addressed in the first research question. The analytic steps and procedure reported below follow this pre-condition order.

**Theoretical framework**

**Teaching behaviour and its measurement**

Teaching behaviour is a central concept in the literature of teaching, teacher effectiveness, learning environments, and motivation. In general, research has consistently documented the significant role of teaching behaviour for student learning and outcomes (e.g., Chapman et al., 2016; Creemers & Kyriakides, 2008; Hattie, 2012; Hattie & Clinton, 2008; Scheerens & Bosker, 1997). To illustrate effective teaching behaviour for research and practice purposes, measurement plays a substantial role. van de Grift (2014) reported various early observational studies to study teaching quality arising from as early as 1961. He argued that these studies are mostly small in scale, focus on narrow teacher training and observation instruments, are not grounded within a solid research framework on educational effectiveness, and are insufficiently attentive to the reliability and validity aspects in an international context (van de Grift, 2014). Regardless of these limitations, the early work of observation studies paved the way for more refined and robust observation studies on teaching quality.

To date, some popular and well-known observation measures of effective teaching grounded in strong teaching effectiveness and motivational frameworks exist. These include the dynamic system model (Creemers & Kyriakides, 2008), the Framework for Teaching (FFT; Danielson, 2013), the Classroom Assessment Scoring System (CLASS; Pianta & Hamre, 2009), the three instructional dimensions model (Klieme et al., 2009), the International System for Teacher Observation and Feedback (ISTOF; Muijs et al., 2018), and the International Comparative Analysis of Learning and Teaching (ICALT; van de Grift, 2014). The mentioned observation instruments vary in terms of structure and their main underlying models and conceptualisations. However, the instruments share overlapping concepts and characteristics that are recognised as effective teaching behaviour in teacher effectiveness research (see Maulana et al., 2014; van de Grift et al., 2017).

In the current study, we opted for the ICALT instrument for measuring teaching behaviour for the following reasons. First, the instrument is strongly grounded in the evidence-based teacher effectiveness research covering important observable domains of teaching behaviour. Second, the instrument is economic and user friendly due to a
manageable number of items, comprehensible language use, and simple format, which makes it attractive for researchers and teachers internationally. Third, prior research indicated that the instrument is relevant for use in primary (van de Grift, 2014) and secondary (Maulana et al., 2017) education, adding to the external validity in terms of educational level. Furthermore, the instrument is also applicable for use in several European countries (van de Grift et al., 2014) as well as in contrasting cultural contexts like South Korea (Maulana et al., 2019; van de Grift et al., 2017), providing preliminary evidence that the instrument can be translated across national borders. Finally, the instrument has been continuously used in the Netherlands as a diagnostic tool for teacher professional development of both pre-service and in-service teachers (Bell et al., 2019), suggesting that the instrument is highly useful for both research and practice, and it is relatively easy to train observers.

**Teaching behaviour: the ICALT observation measure**

ICALT\(^1\) was initially developed in 2002 through cooperation between several European inspectorates to measure generic teaching behaviour. Five years later, van de Grift (2007) reported the potential use of the measure for comparing teaching behaviour across several primary schools in Europe. Measurement invariance of ICALT was not covered by this study. In 2014, the first attempt to validate ICALT across primary schools in Europe employing a linear confirmatory factor analysis (CFA) was conducted (van de Grift, 2014). In 2017, ICALT was validated for use in Dutch secondary education (Maulana et al., 2017). The measure was further validated for use in South Korean (van de Grift et al., 2017) and Indonesian (Irnidayanti et al., 2020) secondary education. Hence, ICALT has been used in a number of studies involving solo and duo countries. No studies involving secondary education in more than two countries on ICALT are currently present.

On the basis of reviews of evidence-based teacher effectiveness research, van de Grift (2007) identified three indirectly observable and six directly observable domains of effective teaching behaviours. The three indirectly observable domains are setting minimum goals, sufficient learning and instruction time, and monitoring students’ learning process. These domains are difficult to observe during a typical lesson (Wubbels et al., 2006) and require a longitudinal teacher observation or different methods to measure. The focus of the current study is on observable domains of teaching behaviour as covered by ICALT. Therefore, the six observable domains are discussed further: creating a safe and stimulating learning climate, providing efficient classroom management, displaying clarity of instruction, activating teaching, adapting instruction to students’ learning needs (differentiated instruction), and teaching the students learning strategies. The six observable domains of teaching quality are in accordance with earlier empirical findings of effective teaching behaviour (e.g., Danielson, 2013; Klieme et al., 2009; Ko & Sammons, 2013;Muijs et al., 2018; Pianta & Hamre, 2009).

The first domain is creating a safe and stimulating learning climate. This includes the provision of a positive climate for learning to take place such as promoting inclusive respect for all students in the classroom and increasing self-confidence of students (Cornelius-White, 2007; Hattie & Clinton, 2008; Smith et al., 2008). The second domain, providing efficient classroom management, covers indicators of lesson organisations such as
good lesson preparation, time management, effective lesson transition management, maximising task time, and efficient handling of student’s misconduct (Marzano, 2003; Opdenakker & Minnaert, 2011; Maulana et al., 2012; M. C. Wang et al., 1995; Yair, 2000). The third domain, clarity of instruction, includes aspects of instructional quality. Indicators of teaching behaviour in this domain involve a clear lesson structure, good interchange of explanations, structured assignments of individual and group works (Kindsvatter et al., 1988; Mortimore et al., 1988; Rosenshine, 1980), and regularly checking of students’ understanding of learning material (Hattie & Clinton, 2008; Kame‘enui & Carnine, 1998; Pearson & Fielding, 1991; Rosenshine & Meister, 1997; Smith et al., 2008).

The fourth domain, activating teaching, deals with behaviour to maximise learning outcomes including active learning, intensive instructions, and avoidance of excessive work seats (Hampton & Reiser, 2004; Lang & Kersting, 2007). Other indicators involve activation of prior knowledge and making use of “advance organisers” (Nunes & Bryant, 1996; Pressley et al., 1992). The fifth domain, differentiated instruction, covers aspects of teacher behaviour related to attending to the diverse backgrounds and personalities of students. Differentiation is defined as a philosophy of teaching in which teachers proactively modify curricula, teaching methods, resources, learning activities, or requirements for student products to better meet students’ learning needs (Tomlinson et al., 2003). The objective is to optimise the learning potential of each student (Tomlinson et al., 2003). Research showed that differentiated teaching leads to improved student outcomes (Koeze, 2007; Reis et al., 2011). Indicators of teachers’ teaching quality are associated with extra time and additional instructions, pre-teaching and re-teaching, and effective teaching methods (Houtveen et al., 1999; Kindsvatter et al., 1988; Lundberg & Linnakyla, 1993; Pearson & Fielding, 1991; Sijtstra, 1997).

The last domain of observable teaching behaviour is teaching learning strategies, which entails teachers’ support for students regarding the use of metacognitive strategies in their classroom teaching. These strategies provide a framework to help students achieve academic success (Carnine et al., 1988). An example is scaffolding, which is a form of temporary support provided by teachers (or by peers) that functions as a bridge between students’ existing and desired skills. Teachers who provide their students with learning strategies have a significant impact on their learning performance (Hattie & Clinton, 2008; Houtveen & van de Grift, 2007a; Rosenshine & Stevens, 1986; Slavin, 1996; Smith et al., 2008).

Past studies indicate that the six domains of teaching behaviour follow a systematic difficulty level, ranging from more basic (learning climate, classroom organisation, clarity of instruction) to more complex skills (activating teaching, teaching learning strategies, differentiated instruction; e.g., van de Grift et al., 2014; van der Lans et al., 2018).

**Teaching behaviour: differences across countries**

Adamson (2012) identified three approaches to comparative educational research. The first is a classical approach at the country or system level emerging from the early 19th century, which is typically manifested in the form of a two-location study. Another approach treats one location as a reference of analysis and then makes comparisons with other locations. A third approach uses several locations with equal treatment in each location. The current study fits closely to the third approach. Notably, typical
cross-country comparison studies on teaching practises have relied mostly on self-reports using student and/or teacher surveys, while we use observations.

The International School Effectiveness Research Project (ISERP) has been recognised as a large-scale study on school and teacher effectiveness across countries, involving Europe, North-America, Pacific Countries, Canada, and Australia (Reynolds et al., 2002). The study indicates that most factors known from national school and teacher effectiveness research are relevant to different national contexts. However, this project also revealed that the main instrument (The Virgilio Teacher Behaviour Inventory, VITB) has limited external validity, due to the fact that the study only included nine developed countries with strong school and teacher effectiveness traditions (Reynolds et al., 2002). This study has paved the way for more detailed explorations of teaching behaviour across national contexts using a uniform and cross-nationally valid observation measure. Subsequently, the International System for Teacher Observation and Feedback (ISTOF) was created as a product of what was learned from the ISERP project. The ISTOF project is a lengthy and ambitious project involving 20 countries (Reynolds et al., 2002). The psychometric quality of ISTOF has been reported in single-country studies. Muijs and colleagues (2018) reported that ISTOF has satisfactory reliability across single-country studies, but the 7-factor structure was not always found. In addition, the factor invariance of ISTOF across countries remains unknown (Muijs et al., 2018), which makes cross-country comparison using the instrument difficult to employ. Hence, there is still a gap in the body of knowledge regarding the need for developing a cross-nationally valid instrument to measure teaching behaviour.

Past studies using ICALT suggest that the instrument seems to be promising for comparing teaching behaviour across countries. van de Grift (2014) reported that ICALT is sufficiently invariant across primary schools in several European countries, although full scalar invariance was not found for some teaching behaviour domains. From secondary education contexts, van de Grift et al. (2017) reported measurement invariance of ICALT across Dutch and South Korean samples. They found that Dutch teachers were rated more positively on learning climates and activating teaching compared to South Korean teachers. On the contrary, Dutch teachers were rated less positively on teaching learning strategies than South Korean teachers. A more recent study focusing on differentiated instruction using ICALT showed that differentiated instruction was rated more positively in South Korea than in the Netherlands. Regardless of this difference, the quality level of differentiated instruction in both countries was observed as being low (Maulana et al., 2019).

**Context of the current study**

Comparing teaching behaviour across various contrasting national settings is a very interesting, yet demanding, endeavour. In this study, teaching behaviour of secondary school teachers in the Netherlands, South Korea, Indonesia, South Africa, Hong Kong-China, and Pakistan were investigated. In terms of student performance as documented by PISA and TIMSS studies, the Netherlands, South Korea, and Hong Kong have been generally ranked as the world’s top educational systems. On the other hand, Indonesia has been frequently ranked as one of the lowest performers. Based on TIMMS 2015
results, South Africa is also on the bottom of the ranking. Pakistan has not taken part in the PISA and TIMSS studies yet.

**The Netherlands**

International comparisons in primary and secondary education showed students attending Dutch schools perform above average, comparable to other high-performing European and Asian educational systems (Martin et al., 2016a, 2016b; Mullis et al., 2016, 2017; OECD, 2016). The majority of teenagers master at least the basic skills in reading, mathematics, and science. The Dutch educational system is highly tracked (i.e., students are split by ability in a large number of different educational tracks from the age of 12), does not apply a national curriculum, shares national educational standards, and gives extensive autonomy to schools and teachers (OECD, 2014, 2016). The high level of decentralisation is balanced by a strong school inspection mechanism and a national examination system. 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 (Inspectie van het Onderwijs, 2018; OECD, 2016).

**South Korea**

The South Korean educational system is among the top-performing systems showing an excellent performance compared to most other countries in PISA and TIMSS (Martin et al., 2016a, 2016b; OECD, 2016). This remarkable educational development contributed to the boost of human capital and economic growth (Akiba & Han, 2007). South Korea’s performance reveals a low percentage of underachieving students, and high percentages of excellent students. Tracking starts at the age of 14, which is the same as the OECD average, and grade repetition is rare (OECD, 2016).

High academic achievement is greatly prized in the South Korean society (OECD, 2016). One of the major learning resources in South Korean classes is government-endorsed textbooks, and ICT is used as well (Heo et al., 2018). The South Korean system greatly emphasises teaching quality and ongoing development in the teaching profession. Teachers are recruited from the top graduates, with strong financial and social incentives: social recognition as well as opportunities for career advancement and beneficial occupational conditions (Heo et al., 2018; Kang & Hong, 2008; OECD, 2016). In general, education in South Korea is more teacher centred, although since 2003 new policies regarding the “7th National Curriculum” have been implemented to improve student centredness and student autonomy (Kim, 2003).

**Indonesia**

Indonesia Education Law Number 20 of 2003 states that all citizens have the right to quality education. The central and local governments provide funds to support free basic education. The duration of compulsory education was increased from 9 to 12 years in 2016. Despite the diversity it has that stretches from the western part of Sumatra to the eastern part of Papua, with different landscapes, cultures, religions,
ethnicities, and languages, Indonesia is united in prioritising education. The average education spending has increased significantly each year. In January 2017, the World Bank showed that Indonesia spends 20.6% of the state budget on education (Fasih et al., 2018).

On the basis of TIMSS and PISA, Indonesia is positioned consistently among the lowest performing educational systems. There are many factors that contribute to the low quality of education in Indonesia. The quality of teachers plays a significant role herein. Although teachers in Indonesia should take a certification programme to improve their teaching, the teaching certification is not correlated with students’ learning outcomes (World Bank, 2015). The teaching certification does not require the teachers to implement or demonstrate their knowledge and skills in the classroom. Most teachers employ a teacher-centred approach. Other issues including teacher motivation, teacher selection, and initial teacher training programmes are mentioned as factors explaining the low quality of education in Indonesia (de Ree, 2016; Fasih et al., 2018).

**South Africa**

Despite a government budget of approximately 15% spent on education, South African students reveal very low literacy and numeracy levels and rank last in mathematics and science based on TIMSS 2015. Moreover, of 139 participating countries South Africa scored number 137 for overall quality of education (Baller et al., 2016). The reason for the low reading skills could be that South Africa has 11 official languages and students are instructed in a second language, English (Howie et al., 2012; Spaull, 2013). In addition, teachers experience a lack of reading resources (Zimmerman & Smit, 2014). Changes were evident in curricula that strived to ensure access to education for previously disadvantaged students and to accommodate diverse cultures. In six South African universities, only 6% of the curriculum for teacher training and development includes how a teacher should teach a student to read (Taylor et al., 2013).

Two of the issues that impede inclusive education could be insufficient training of teachers in effective teaching such as differentiated instruction (Dalton et al., 2012) and students’ inadequate English proficiency skills (Neeta & Klu, 2013). Low levels of competence in the instruction language English impede South African students’ academic performance (Cheatham et al., 2014). In addition, Grosser and Nel (2013) and Masitsa (2004) argue that the language of students’ instruction is essential not only for communication but also to construct significant content knowledge. Lomofsky and Lazarus (2001) and Holz and Lessing (2002) indicate that the majority of teachers feel unprepared and inadequately trained for including all learners in high-quality teaching including differentiated learning activities.

**Hong Kong-China**

Hong Kong has an effective education system based on its top performance in comparative studies (Martin et al., 2016a, 2016b; Mullis et al., 2016; OECD, 2016). The expenditure on education has increased in the past decades with free comprehensive primary and secondary education in public schools since 2008 and significantly subsidised kindergarten education since 2006 (Education Bureau, 2018). Hong Kong schools have strong autonomy over curricular and personnel matters, but the government and the public also expect a
high degree of educational and professional accountability (Ko et al., 2016). The curriculum is policy driven and standardised by the government at the system level, but the curriculum and its reforms relying on school-based efforts to make changes in daily practices are not effective or sustainable (Lee et al., 2018). Like other Confucian societies, assessment of learning is driven by high-stake examinations and Chinese cultural norms while assessment for student improvement is gradually recognised among teachers despite barriers (G. T. L. Brown, 2008; G. T. L. Brown et al., 2009). The respect and status of teachers have been enhanced with the teacher competencies framework and the continuous professional development programmes implemented in the early 2000s (H. Y. Cheung, 2008; Ko et al., 2016; Ko & Sammons, 2013).

Education reforms in China have shown positive economic and social impacts (Hannum & Park, 2007). In PISA 2015, the performances in mathematics and science of a broader area of China including Beijing, Shanghai, Jiangsu, and Guangdong (B-S-J-G) were more modest compared to previous achievements by Shanghai, but still above the OECD averages (OECD, 2016). Chinese schools have stronger autonomy over resources but stricter governmental controls over curricular and personnel matters. Accountability exists in the system as social expectations of school leadership and teacher professionalism rather than quality assurance mechanism (OECD, 2016).

Apart from the traditional respect towards teachers, teachers are valued when they demonstrate their professionalism through their students’ performances, research, and teaching contests (Chen et al., 2012). Large differences in educational expenditure across provinces and counties have contributed to variations in teacher quality (P. H. Brown & Park, 2002), but reforms are undergone at all levels to increase financial inputs to improve teacher education (Ministry of Education, 2014). Teacher professional development is gradually moving from the traditional lecture-centred, theory-orientated, and experience-led modes to school-based (Gu & Wang, 2006), learning community (Sargent & Hannum, 2009), lesson study (Huang & Bao, 2006), and demonstration lesson modes (OECD, 2011).

**Pakistan**

Pakistan ranks 113th out of 120 countries on the educational performance index (United Nations Educational, Scientific and Cultural Organization [UNESCO], 2015). It spends only 2.1% of its gross domestic product on education. Pakistan mirrors the third largest adult illiteracy in the world. Up to 49.5 million adults are illiterate, from which two thirds is female. Almost half of the young rural women in Pakistan never get a chance to go to school (UNESCO, 2015). Studies indicate that initial teacher education in Asian countries including Pakistan is facing a considerable challenge regarding deficiencies in policy making, goal planning, curriculum development, teacher training, and the adoption of varying types of didactical approaches (Ahmad et al., 2014; Morris & Williamson, 2000; Sharma et al., 2013).

Nearly 90% of all teachers received an initial education, but the results of their teacher training are not reflected in student outcomes (Ali, 2011). Surveys, international agency reports, educational policies, and government’s 5-year plans indicate that the quality of initial teacher education and in-service teacher education programmes in Pakistan are not up to international standards (UNESCO, 2006). Teacher educators are unable to
inline their didactical strategies with the desired objective of initial teacher education. Additionally, there is an absence of lesson planning element in the entire teaching activity in teacher education institutes in Pakistan (Rizvi, 2015).

Method

Sample and procedure

The current study included 2,680 teachers’ data from six countries involving the Netherlands ($N_{teacher} = 524$), Indonesia ($N_{teacher} = 452$), South Africa ($N_{teacher} = 313$), South Korea ($N_{teacher} = 773$), Hong Kong-China ($N_{teacher} = 218$ in which 72 from Hong Kong and 146 from Shenzhen and Guangzhou), and Pakistan ($N_{teacher} = 400$). Because of the small sample size, data from Hong Kong and China were combined together. Of the Dutch teachers, 32.3% taught natural science subjects, 59.9% were female, 67% were inexperienced, 22.9% taught in vocational schools, and all were from public schools. Of the Indonesian teachers, 42.9% taught natural science subjects, 62.2% were female, 20.4% were inexperienced, 8.4% taught in vocational schools, and 76.1% were from public schools. Of the South African teachers, 40.58% taught natural science subjects, 47.3% were female, 24.3% were inexperienced, all were from general (non-vocational) schools, and 2.9% from private schools. Of the South Korean teachers, 35% taught natural science subjects, 65.7% were female, 2.4% were from vocational schools, and 19% from private schools. The sample from Hong Kong-China only included English (57%) and mathematics (43%) lessons. Of the Pakistani teachers, 50.3% were female, 59.8% were inexperienced, 11.1% were from vocational schools, and all were from public schools.

Similar sampling strategies were employed in the six countries. A stratified sampling method was initially applied, which resulted in low response rates. Subsequently, convenience sampling was applied. Invitations to participate followed the common bureaucratic procedure in each country, which typically included agreements between researchers and the local education authority, schools, and teachers. All teachers participated on a voluntary basis. Typical lessons of the participating teachers were observed in the natural classroom settings. Data were collected in different years ranging from 2014 to 2018. The full lesson of each participating teacher was observed once.

Measure

To measure teaching behaviour in the six countries, the ICALT observation instrument was used (van de Grift et al., 2014). The instrument consists of 32 high inferential observable teaching acts, accompanied with 120 low inferential observable teaching activities. The low inference indicators are examples of good practices associated with each high inference item. The 32 high inference items represent the six domains of teaching behaviour discussed in the literature section, which include safe and stimulating educational climate (4 items), efficient classroom management (4 items), clarity of instruction (7 items), activating teaching (7 items), differentiated instruction (4 items), and teaching learning strategies (6 items). Previous research has confirmed the six-factor structure of observed teaching behaviour in Dutch secondary education (Maulana et al., 2017). Observers rated the items on a 4-point scale: mostly weak, more often weak than strong, more often strong than weak, and strong. Because of the skewed
distribution of answers in the sample, the categories mostly weak and more often weak than strong were recoded into weak. This decision was necessary to improve model identification and fit (J. Wang & Wang, 2012). Examples of the domains and the corresponding items are given in Appendix 1.

The original English version of the instrument was used as the source language for the translation and back-translation procedure. The target language of translation included Indonesian, Korean, and Chinese. English was used as the language of instruction in South Africa, Pakistan, and Hong Kong-China. Hence, the English version was used in these countries. The process was done following the guidelines of the International Test Commission (Hambleton, 1994). In each country, the structured translation and back-translation process involved two highly knowledgeable researchers concerning the instrument and the theoretical framework underlying the instrument and two university professors proficient in both English and the target languages. Upon the completion of the procedure, issues and discrepancies were discussed thoroughly and resolved subsequently by the team.

The issues were relatively minor and revolved around word equivalence and the accuracy of word choice. The expert team confirmed the relevance of the six domains of teaching behaviour in their own national contexts, providing evidence for face validity.

**Observer training**

The observer training aims to obtain mutual consensus between observers scoring the same lessons and consensus between the observers and the norm generated by a large number of experts (experienced teachers and educators). In all six countries, identical training in terms of standard, structure, and procedure for using the ICALT observation instrument were applied. Two experts’ trainers from the Netherlands conducted the training in the six countries directly. Due to unavoidable circumstances, however, the training in Pakistan was done via online training. The training in all countries consisted of three phases: preparation, implementation, and evaluation.

In the preparation phase prior to the training day, trainees were instructed to read and study the theoretical framework underlying the instrument carefully. They were also instructed to study the instrument thoroughly. The implementation phase refers to the interactive, in person, training day which lasted for one full day. During this phase, explanations and discussion about the instrument and the corresponding theory were conducted. Afterwards, discussions about how to evaluate teaching practices using the associated scoring rules were held. Furthermore, two videotaped lessons (English and geography) were used by observers respectively to rate the two teachers’ teaching behaviour using the observation instrument. The consensus level of 70% within the group and between the group and the expert norm was set as a sufficient cut-off. Discussions to resolve significant differences and improve consensus were conducted subsequently. Finally, the evaluation phase involved the investigation of rating patterns and significant deviations from the average pattern. A small number of observers who deviated from the average were followed up, and extra guidance was given to this group prior to conducting the observation in the natural classroom settings. Observers failing to meet the minimum consensus were not invited to conduct observations.

A popular method for computing a consensus estimate (of interrater reliability) by using a simple agreement percentage was applied: a product of adding up the number of items
that received identical ratings by all observers and dividing that number by the total number of items rated by observers (Stemler, 2004). The consensus estimates from the observer training were generally sufficient (see Table 1). The within-group consensus among South African observers was rather low (0.63), but the consensus with the expert norm was above satisfactory (0.88). The reverse pattern was found for Pakistan, in which the consensus with the expert norm was 0.63.

Analytic approach

Before conducting measurement invariance testing, the measurement model of the ICALT measure was tested in each country separately using CFA. After the measurement model in each country was confirmed, multigroup confirmatory factor analysis (MGCFA) combining all country data was performed. All analyses were done using MPlus version 8.1 (Muthén & Muthén, 2017). We followed the examples provided by Millsap and Yun-Tein (2004) and Xing and Hall (2015) to estimate the models. Three levels of measurement invariance were tested, respectively. First, configural invariance tests whether the same factor structure can be applied in each country (in all countries all items load on the same factor). Second, metric invariance tests whether factor loadings are equal across countries. Third, scalar invariance tests whether both factor loadings and also the intercepts of ICALT items are equal across countries. Establishing scalar invariance means that we can meaningfully compare the means of the factors across countries (Byrne, 2012).

The common model-data goodness of fit indices for the categorical CFA and MGCFA models include the root-mean-square error of approximation (RMSEA), the comparative fit index (CFI), and the Tucker–Lewis index (TLI), and adhere to common guidelines (i.e., CFI and RMSEA and lower and upper for 90% confidence interval of RMSEA < 0.08; CFI > 0.90; TLI > 0.90) for an acceptable model fit (T. A. Brown, 2015; Hu & Bentler, 1999). A second approach to assess the measurement invariance, especially in the case of small groups, is to test the deterioration of the model fit between the configural, metric, and scalar model. Changes in CFI (ΔCFI), TLI (ΔTLI), and RMSEA (ΔRMSEA) of < 0.01 were applied (G. W. Cheung & Rensvold, 2002).

Results

Reliability of teaching behaviour

Reliabilities of the ICALT measure in all countries are above satisfactory level (Cronbach’s alpha values range between 0.74 and 0.92). The full measurement model of CFA in each

<table>
<thead>
<tr>
<th>Country</th>
<th>Consensus estimates</th>
<th>Consensus estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Within the group</td>
<td>With the expert norm</td>
</tr>
<tr>
<td>Netherlands</td>
<td>71</td>
<td>86</td>
</tr>
<tr>
<td>South Korea</td>
<td>80</td>
<td>88</td>
</tr>
<tr>
<td>Indonesia</td>
<td>74</td>
<td>67</td>
</tr>
<tr>
<td>South Africa</td>
<td>63</td>
<td>88</td>
</tr>
<tr>
<td>Hong Kong-China</td>
<td>77</td>
<td>75</td>
</tr>
<tr>
<td>Pakistan</td>
<td>100</td>
<td>63</td>
</tr>
</tbody>
</table>

Table 1. Consensus estimates of ICALT observer training in the six countries.
country data shows the model fits moderately well in all included countries, except in Hong Kong-China (see Table 2). The modification indices show the model-data fit could be improved by releasing one or several constraints of the model in some countries. We decided not to do this at this point for two reasons. First, we aimed to obtain a cross-culturally valid measurement model of teaching behaviour. Therefore, constraints released should be released in none or all countries to obtain full measurement invariance. Second, the number of cases in each group is sufficient, but small enough to expect the fit to increase when we combine the countries into one model.

For the Hong Kong-China data, however, the acceptable functioning measurement model was not obtained. One of the main problems involved Item 24 (“offers weaker learners extra study and instruction time”; Domain: Differentiated instruction) showing no cell filling in the three response categories, which caused a convergence problem. Furthermore, the Hong Kong-China data are relatively small compared to data from the other countries. A more detailed inspection of the model when leaving Item 24 out showed the same factor structure as in the other countries, which can be indicative of measurement invariance for Hong Kong-China. However, an acceptable model fit was not reached even after deleting this item. In the next analysis step, the Hong Kong-China data were excluded.

Factor loadings for the six domains in the five countries are above the common 0.40 threshold (see Table 3), indicating that all items load on their corresponding domains sufficiently. Since we did not establish cross-cultural invariance at this stage, it is not recommended to compare the loadings between countries, only within countries.

**Measurement invariance of teaching behaviour across countries**

Results of categorical MGCFA with the five country data revealed no convergence, although the six-factor structure was confirmed in the six countries separately. A closer inspection of the issue indicated that the Pakistan data have no full filling on Item 28 (“stimulates the use of control activities”; Domain: Teaching learning strategies). To solve this issue, two decisions were made: (a) checking the model without the Pakistan data, (b) leaving Item 28 out in the full five country analyses.

Excluding the Pakistan data showed that the measurement invariance holds in the remaining four countries (see Table 4). The configural model is just below the minimum threshold. The decrease in fit when restraining all the factor loadings to be equal across the four countries is quite small, indicating that a reasonably good fit is evident compared
to the configural model. Constraining intercepts across the four countries for all latent variables resulted in a smaller decrease in model-data fit. This indicates that scalar equivalence across the four countries is evident.

Including Pakistan, and excluding Item 28 in the five countries leads to a moderate scalar equivalent model with RMSEA = 0.098, CFI = 0.0903, and TLI = 0.971; based on the decision rule that RMSEA < 0.08, CFI > 0.90, and TLI > 0.90 for an acceptable model fit for the configural model (T. A. Brown, 2015; Hu & Bentler, 1999). Nevertheless, the
configural model indicates some problems, which compromises the interpretation of the metric and scalar models. This indicates that although the model fits better in the other four countries including Item 28, the Pakistan data results bear some comparability with the other countries on the ICALT measurement. Because our aim is to confirm the original 32-items ICALT factor structure involving various national samples instead of selecting “comparable” items that fit for all countries, results from the four countries meeting the full scalar invariance are discussed further and Pakistan is excluded. In Table 4, we present the results of the categorical multigroup confirmatory factor analysis. Although the configural invariant model shows insufficient invariance based on the RMSEA and CFI, the TLI measure is well above the standard. On the basis of the decision rules of Hu and Bentler (1999), we would thus reject this model. However, G. W. Cheung and Rensvold (2002) stated that using only the absolute fit measures is not the best approach to follow. Using changes in fit indices is a better approach because it is independent of sample size and model complexity. We therefore also look at changes in the fit indices, which should be < 0.01 to show invariance. We find this change in fit between the configural/metric, and metric/scalar model for all three fit indices. This strengthens the evidence that the ICALT observation instrument shows measurement invariance in the countries included in this study. Standardised factor loadings for each country in the scalar-invariant model can be seen in Table 5.

Having reached full scalar invariance with a reasonably fitting model with absolute fit (RMSEA) and a good comparative fit (CFI/TLI), comparing latent means in the four countries is justifiable (see Table 6).

**Table 5.** Standardised factor loadings for each country in the scalar-equivalent model.

<table>
<thead>
<tr>
<th>Domains and items</th>
<th>Factor loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>The Netherlands</strong></td>
<td></td>
</tr>
<tr>
<td>Learning Climate (1–4)</td>
<td>0.872 0.736 0.914 0.779</td>
</tr>
<tr>
<td>Classroom Management (5–8)</td>
<td>0.764 0.763 0.725 0.669</td>
</tr>
<tr>
<td>Clarity of Instruction (9–15)</td>
<td>0.714 0.690 0.740 0.786 0.709 0.704 0.655</td>
</tr>
<tr>
<td>Activating Teaching (16–22)</td>
<td>0.616 0.615 0.716 0.770 0.705 0.652 0.440</td>
</tr>
<tr>
<td>Differentiated Instruction (23–26)</td>
<td>0.821 0.739 0.702 0.887</td>
</tr>
<tr>
<td>Learning Strategies (27–32)</td>
<td>0.75 0.805 0.746 0.714 0.709 0.751</td>
</tr>
<tr>
<td><strong>Indonesia</strong></td>
<td></td>
</tr>
<tr>
<td>Learning Climate (1–4)</td>
<td>0.643 0.767 0.648 0.839</td>
</tr>
<tr>
<td>Classroom Management (5–8)</td>
<td>0.825 0.844 0.863 0.788</td>
</tr>
<tr>
<td>Clarity of Instruction (9–10)</td>
<td>0.781 0.628 0.835 0.631 0.820 0.846 0.729</td>
</tr>
<tr>
<td>Activating Teaching (16–22)</td>
<td>0.705 0.709 0.655 0.625 0.653 0.672 0.769</td>
</tr>
<tr>
<td>Differentiated Instruction (23–26)</td>
<td>0.874 0.795 0.802 0.889</td>
</tr>
<tr>
<td>Learning Strategies (27–32)</td>
<td>0.737 0.761 0.774 0.777 0.796 0.760</td>
</tr>
<tr>
<td><strong>South Africa</strong></td>
<td></td>
</tr>
<tr>
<td>Learning Climate (1–4)</td>
<td>0.964 0.971 0.931 0.943</td>
</tr>
<tr>
<td>Classroom Management (5–8)</td>
<td>0.974 0.940 0.947 0.917</td>
</tr>
<tr>
<td>Clarity of Instruction (9–15)</td>
<td>0.870 0.818 0.910 0.836 0.849 0.876 0.828</td>
</tr>
<tr>
<td>Activating Teaching (16–22)</td>
<td>0.877 0.890 0.882 0.884 0.818 0.867 0.822</td>
</tr>
<tr>
<td>Differentiated Instruction (23–26)</td>
<td>0.900 0.939 0.942 0.928</td>
</tr>
<tr>
<td>Learning Strategies (27–32)</td>
<td>0.923 0.949 0.917 0.928 0.905</td>
</tr>
<tr>
<td><strong>South Korea</strong></td>
<td></td>
</tr>
<tr>
<td>Learning Climate (1–4)</td>
<td>0.808 0.750 0.902 0.886</td>
</tr>
<tr>
<td>Classroom Management (5–8)</td>
<td>0.795 0.854 0.839 0.772</td>
</tr>
<tr>
<td>Clarity of Instruction (9–15)</td>
<td>0.787 0.747 0.782 0.806 0.772 0.815 0.791</td>
</tr>
<tr>
<td>Activating Teaching (16–22)</td>
<td>0.732 0.709 0.828 0.800 0.706 0.740 0.742</td>
</tr>
<tr>
<td>Differentiated Instruction (23–26)</td>
<td>0.858 0.767 0.859 0.856</td>
</tr>
<tr>
<td>Learning Strategies (27–32)</td>
<td>0.786 0.832 0.846 0.799 0.753 0.812</td>
</tr>
</tbody>
</table>
Teaching behaviour across countries

Using the Netherlands as a reference category, Indonesian teachers were rated lower on four of the six domains, but higher on teaching learning strategies ($p < 0.001$). No significant difference was found for differentiated instruction. Furthermore, South Korean teachers were rated higher on all domains ($p < 0.001$), except for learning climate. South African teachers scored lower on learning climate, classroom organisation, and clarity of instruction ($p < 0.01$), but higher on activating teaching, differentiated instruction, and teaching learning strategies ($p < 0.01$; see Table 6, Figure 1).

Table 6. Comparison of latent means of the scalar invariance MGCFA model.

<table>
<thead>
<tr>
<th>Domains</th>
<th>South Korea</th>
<th>Indonesia</th>
<th>South Africa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning climate</td>
<td>0.005</td>
<td>-0.902***</td>
<td>-0.566***</td>
</tr>
<tr>
<td>Classroom organisation</td>
<td>0.249***</td>
<td>-0.542***</td>
<td>-0.380***</td>
</tr>
<tr>
<td>Clarity of instruction</td>
<td>0.178***</td>
<td>-0.562***</td>
<td>-0.221**</td>
</tr>
<tr>
<td>Activating teaching</td>
<td>0.568***</td>
<td>-0.426***</td>
<td>0.232**</td>
</tr>
<tr>
<td>Differentiated instruction</td>
<td>1.127***</td>
<td>-0.113</td>
<td>0.904***</td>
</tr>
<tr>
<td>Teaching learning strategy</td>
<td>1.056***</td>
<td>0.362***</td>
<td>1.233***</td>
</tr>
</tbody>
</table>

Note: The Netherlands is the reference category (its value is set to 0). MGCFA = multigroup confirmatory factor analysis. *$p < 0.05$. **$p<0.01$. ***$p < 0.001$.

Figure 1. Latent means on the six domains compared to the Netherlands (the Netherlands is the reference category).
Of all teaching behaviour domains, differentiated instruction was consistently ranked as the lowest in the four countries (see Figure 2). Compared to South Korea and South Africa, this particular domain was rated substantially lower in the Netherlands and Indonesia.

Conclusions and discussion

The main aim of the current study was to test whether an observation instrument measuring effective teaching behaviour is invariant across national contexts. Subsequently, it aims to compare teaching behaviour across national contexts using a uniform, and identical, instrument. The current knowledge on measurement invariance of teaching behaviour measures across cultures is still in development. The current study is one of the first to address this important subject by testing the measurement model and factor structure of the observation measure ICALT (van de Grift et al., 2014) in six contrasting national contexts: the Netherlands, South Korea, Indonesia, South Africa, Hong Kong-China, and Pakistan.

Main findings

Our first question was: To what extent is the use of an observation instrument measuring effective teaching behaviour in secondary education invariant in six national samples (the Netherlands, South Korea, Indonesia, South Africa, Hong Kong-China, and Pakistan)? We found that the six-factor structure of teaching behaviour, as measured by ICALT, is confirmed in the included countries, except in Hong Kong-China. This result provides further empirical evidence supporting the reported face validity by the national experts.
about the relevance of the six teaching behaviour domains in their own national contexts. The poor fit of the six-factor model for Hong Kong-China does not necessarily mean that the teaching behaviour constructs are not applicable in the country. A closer inspection of the data distribution and response patterns indicated that Item 28 ("offers weaker learners extra study and instruction time") has no cell filling in Hong Kong-China. It is possible that the answer on Item 24 signifies something else or shows observer response bias in the observations compared to the other countries. Earlier research showed that Asian students in an international degree programme interpreted and answered questions differently compared to their European peers on an educational success measurement model (Jansen et al., 2013). In addition, the Hong Kong-China sample included in this study is relatively small compared to data from the other countries. Due to the sample insufficiency, it is possible that the current sample does not cover a representative range of teachers measured by this particular item. Insufficient sample size could affect statistical power especially when applied to a complex estimation model like CFA, given the relatively large number of ICALT items.

Although the six domains of teaching behaviour are empirically evident in the five national contexts, including Pakistan in the measurement invariance tests yielded estimation errors. The Pakistan data have no full filling on Item 28 ("stimulates the use of control activities"). A closer look at the data and response patterns indicates that certain items are skewed in a different direction than the other country data, which might cause the estimation problem. The skewness of the data distribution might indicate that the current sample does not cover the representative range of Pakistani teachers or observer response bias. Nevertheless, the findings of the current study suggest that the six domains of teaching behaviour are relevant to the Pakistani context as well, but evidence to support whether teaching behaviour of Pakistani teachers can be compared with other national contexts is limited and requires further investigations. Establishing measurement invariance of teaching behaviour measures is particularly important for Pakistan because the country has been facing challenges to help teachers to meet the minimum quality standards (Ali, 2011; UNESCO, 2006). Subsequently, the country could learn from best teaching practices in other countries with better teaching quality to improve the country's teacher performance for better student outcomes. As educational effectiveness research reveals, teachers and schools matter most for underprivileged and/or initially low-achieving students (Kyriakides, 2004; Scheerens & Bosker, 1997; van Laar et al., 2017).

The second question was: What are the differences and similarities regarding the six domains of effective teaching behaviour across the six national samples? The current study confirms that the teaching behaviour measure is invariant across the Netherlands, South Korea, Indonesia, and South Africa. Hence, comparing teaching behaviour in these four national contexts is deemed acceptable. The four countries can be contrasted in terms of educational systems and student performance. South Korea and the Netherlands have been positioned in the top educational systems and academic performance listing. In contrast, Indonesia and South Africa have been placed in the bottom of the listing (Martin et al., 2016a, 2016b; OECD, 2016). South Korean teachers were rated higher in all domains except learning climate compared to Dutch teachers. South Korean teachers were also rated higher in all domains except teaching learning strategies compared to South African and Indonesian teachers.
Higher ratings on most of the teaching behaviour domains for South Korean teachers compared to Dutch, South African, and Indonesian teachers might be related to several effective teaching supporting factors including how teachers in the country are recruited, how they value learning, and how they are supported professionally. The country recruits teachers from the top graduates, and is able to make the teaching profession a highly respected and secure occupation in society (Heo et al., 2018, Kang & Hong, 2008; OECD, 2016). Learning is greatly valued in the Korean society (OECD, 2016), and the government provides ongoing professional development support to facilitate learning including major textbooks and contemporary ICT facilities (Heo et al., 2018). The mentioned supporting factors may lead to improved teaching behaviour, which may further be reflected in the country’s top student performance (Martin et al., 2016a, 2016b). Although this line of reasoning sounds logical, future research should prove whether this reasoning is empirically validated.

Higher ratings on more basic skills for Dutch teachers compared to South African and Indonesian teachers are in line with the report of the Dutch Inspectorate of Education (Inspectie van het Onderwijs, 2018) and that of the OECD (2016) mentioning that the quality of Dutch teachers is generally high with the large majority mastering the basic teaching skills well. Improving more complex teaching skills such as teaching learning strategies and differentiated instruction has been emphasised in research and in the policy agenda of Dutch education (Maulana et al., 2019).

South African teachers were rated the highest in teaching learning strategies. Compared to Dutch teachers, South African teachers were rated higher on activating teaching, differentiated instruction, and teaching leaning strategies, but were rated lower on learning climate, classroom organisation, and clarity of instruction. The overall quality of education in the country is low (Baller et al., 2016). The majority of teachers felt insufficiently prepared and lack skills for including all students in high-quality teaching including differentiated instruction (de Jager, 2013; Holz & Lessing, 2002; Lomofsky & Lazarus, 2001). Although why South African teachers scored higher in more complex skills than Dutch teachers remains unclear. One possible explanation can be linked to teacher training conditions. Although teachers are well trained, the rather poor infrastructure in schools generally does not allow them to apply the skills they have developed during training. Large class sizes (as large as 90 students in a class), English as a second language (ESL) tuition, insufficient resources, lack of internet access, not enough schools and classrooms, consistent power failure, lack of parental involvement, a lengthy curriculum, and protest actions for free education causing absenteeism from schools may contribute to explaining teaching quality in South Africa.

Of all four countries (the Netherlands, South Korea, Indonesia, South Africa), Indonesian teachers were generally rated lower in all domains, except in teaching learning strategies. Teaching learning strategies of Indonesian teachers was rated slightly higher compared to that of Dutch teachers. This finding is consistent with the quality of education in Indonesia as among the lowest performing educational systems (OECD, 2016). Although the government has employed teacher certification programmes for all in-service teachers to improve teaching quality, this certification does not correlate with students’ learning outcomes (World Bank, 2015).

Of all domains, differentiated instruction was rated the lowest in the four countries (see Figure 2). This finding is in line with previous studies indicating that differentiated
instruction is a complex skill to be practised by teachers (Maulana et al., 2014; van der Lans et al., 2018). This domain was rated as insufficient in the Netherlands and Indonesia, and sufficient in South Korea and South Africa. Even in a high-performing country like South Korea, differentiated instruction remains a difficult skill to implement.

**Implications**

Of the six included countries, we have evidence that the teaching behaviour measure is invariant in four countries (the Netherlands, South Korea, Indonesia, South Africa), even when using complex and strict approaches to testing factor structure and measurement invariance such as categorical CFA and categorical MGCFA without excluding any items (full confirmatory and invariant analyses). Hence, the current study suggests that efforts to compare teaching behaviour using an internationally valid instrument across national borders seem to be promising. Cross-national comparison in teaching behaviour is meaningful if the measurement invariance has been established. As Noah (1984) noted:

Properly done, comparative education can deepen understanding of our own education and society; it can be of assistance to policymakers and administrators; and it can form a most valuable part of the education of teachers. Expressed another way, comparative education can help us understand better our own past, locate ourselves more exactly in the present, and discern a little more clearly what our educational future may be. (p. 551)

Subsequently, the four countries could learn from each other regarding malleable factors to improve teaching behaviour for better student outcomes. In order for this to be successful, productive cooperation between teachers, schools, researchers, and policy makers in the four countries should be established. Despite the promising results, establishing factor structure and measurement invariance of a teaching behaviour measure in broader national contexts remains a big challenge. The current study shows that measurement invariance in Hong Kong-China and Pakistan could not be established yet due to sample-related issues. Attempts to increase more variations in these two national contexts are recommended. Finally, differentiation has been recognised in the literature and policy agenda as a highly important skill for teachers to master (Smale-Jacobse et al., 2019; Tomlinson et al., 2003). The fact that differentiated instruction was found to be the lowest in the Netherlands, South Korea, South Africa, and Indonesia implies that the countries should invest in developing this particular domain further.

**Limitations and future directions**

Despite its strengths, the current study is subject to some limitations. The convenience sampling method makes it unadvisable to generalise the findings to the country level. Readers should view findings of the current study as indicative instead of definitive until studies with broader, more representative, and randomly selected samples are available. Although the overall samples are large, comprising six countries, the proportion of higher level data (i.e., school level) per country is relatively limited, which makes it difficult to apply multilevel MGCFA. Future research should aim to increase the number of higher levels to take the possible effect of multilevel data structure into account in the investigation of measurement invariance. Furthermore, although
the observer training in all the countries was done by the same expert trainers using the same standards, the actual data collection in each country was done by the local observers. Although the local training was successful, as indicated by face validity and consensus estimates, and the observation data were reliable and valid, we acknowledge the vulnerability of cultural influences of local observers in conducting actual observations. Future research should attempt to employ cross-observer designs in all country combinations to minimise cultural observation biases. In addition, future research should attempt to investigate non-invariance more specifically in each country on potential sources of invariance including raters, teacher characteristics, and contextual characteristics, taking into account the (im)balanced proportion of the mentioned variables.

The training in Pakistan was done with an online training instead of employing a face-to-face approach like in the other five countries due to some inevitable challenges. Although the within-group consensus estimate was considerably high, the consensus estimate with the expert norm was rather low. This issue may point to different effects of physical and environmental administration of trainings (van de Vijver & Tanzer, 2004), which could potentially explain the rather low consensus with the expert norm. When circumstances allow, future trainings should be done identically across contexts. Finally, unlike the other five countries, observers in Hong Kong-China rated video-taped lessons instead of observing lessons in the authentic settings. The observers in this country were also not experienced teachers. Perhaps this may partially explain the data response patterns in this country, which resulted in little variations in the scores to more difficult items such as was the case with Item 24 measuring an aspect of differentiated instruction. Because differentiated instruction is complex to implement (Maulana et al., 2019), measuring actual differentiation practices may require experienced observers to observe lessons in the authentic setting.

Conducting cross-country comparison studies involving authentic classroom observations is beneficial and innovative. However, it is a highly challenging endeavour. Financial resources have been mentioned frequently as one of the main challenges. Furthermore, complex bureaucratic systems in certain countries impede the possibility of randomised participation in the research. In many countries, classroom observations are not a common practice. Although many Western settings (i.e., the Netherlands, the USA, the UK) value classroom observations highly as a means for improving teaching quality through national inspectorates (van de Grift et al., 2014) and for teacher professional development aims (Helms-Lorenz et al., 2016), other countries value classroom observations less for various reasons including, but not limited to, cultural and personal norms. Finding solutions to solve challenges and extending classroom observation studies in wider cross-country comparison perspectives will likely create bigger impacts on the field of teacher effectiveness and teacher improvement.

Note

1. The instrument was initially named Quality of Teaching (QoT; Ko & Sammons, 2013).

Acknowledgements

The University of Groningen led the research project. The authors from partner universities contributed equally to this work. We would like to specially thank Wim van de Grift (professor emeritus), who provided continuous support during the inception of the current study. We are also indebted
to Peter Moorer for his assistance with the data management, and to Anna Verkade, Geke Schuurman, and Carla Griep for their valuable contributions as international trainers. The biggest thanks goes to all teachers, schools, and observers participating in this large-scale study in all six countries.

**Disclosure statement**

No potential conflict of interest was reported by the authors.

**Funding**

This work was supported by the Dutch scientific funding agency (NRO) under Grant number 405-15-732; the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea under Grant number NRF-2017S1A5A2A03067650; and the Directorate General of Higher Education of Indonesia under Grant number 04/SP2H/DRPM/LPPM-UNJ/III/2019.

**Notes on contributors**

*Ridwan Maulana* is an associate professor at the Department of Teacher Education, University of Groningen, the Netherlands. His major research interests include teaching and teacher education, factors influencing effective teaching, methods associated with the measurement of teaching, longitudinal research, cross-country comparisons, effects of teaching behaviour on students’ motivation and engagement, and teacher professional development. He has been involved in various teacher professional development projects including the Dutch induction programme and school–university-based partnership. He is currently a project leader of an international project on teaching quality (ICALT3/Differentiation) involving countries from Europe, Asia, and Africa. He has been serving as European Editor of *Learning Environments Research* and as a reviewer of various peer-reviewed international journals, as treasurer of the Learning Environments SIG of the American Educational Research Association, and as chair of the Ethical Commission of the Teacher Education department in his university.

*Stéfanie André* is an assistant professor at the Department of Sociology at Radboud University and Assistant National Coordinator of the European Social Survey in the Netherlands. Furthermore, she worked as a post-doc at the Department of Teacher Education at the University of Groningen. Her research interests are in gender roles, part-time work, education, housing, and social inequality. She defended her PhD thesis *The Two Sides of Homeownership, Security and Insecurity* at Tilburg University in 2017. She holds a Research Master’s degree in Social-Cultural Science and a Master’s degree in Comparative Political Science, both from Radboud University Nijmegen. She has published in *Social Science Research*, *Comparative Politics*, and the *International Journal of Comparative Sociology*.

*Michelle Helms-Lorenz* is an associate professor at the Department of Teacher Education, University of Groningen, the Netherlands. Her research interest covers cultural specificity versus universality (of behaviour and psychological processes). This interest was fed by the cultural diversity in South Africa, where she was born and raised. Michelle’s second passion is education, the bumpy road towards development. Her research interests include teaching skills and well-being of beginning and pre-service teachers and effective interventions to promote their professional growth and retention. She has been a project leader of various teacher professional development projects including the national Dutch Induction programme and school–university-based partnerships.

*James Ko* is an assistant professor at the Department of Education Policy and Leadership, The Education University of Hong Kong (EduHK). He is also a research fellow at the Joseph Lau Luen Hung Charitable Trust Asia Pacific Centre for Leadership and Change of EduHK. He has been the grant holder of various projects funded by Hong Kong as well as the UK. His research focuses on issues related to educational effectiveness, educational leadership, teacher evaluation, and teacher professional development. He has published in international journals including *School Effectiveness and School Improvement*, *School Leadership and Management*, the *International Journal of Educational*

**Seyeoung Chun** is a professor at the Department of Education, Chungnam National University, South Korea. He received his PhD degree from Seoul National University. His research focus is on educational policy. He has served in the national education community by holding several key positions, such as the Secretary of Education to the President and the CEO of the Korean Educational Research and Information Service (KERIS). He is the founder and current president of Smart Education Society with more than 3,000 members.

**Abid Shahzad** is an assistant professor at the Department of Education, The Islamia University of Bahawalpur, Pakistan. His major research interests include teaching and teacher education, didactical strategies, teachers’ cognitions, and teachers’ professional development. He has been involved in various teacher professional development projects including the BOF research project (funded by Ghent University Belgium) on teacher education in Pakistan. He is currently the Principal Investigator (PI) from Pakistan for a project on teaching quality (ICALT3/Differentiation) involving countries from Europe, Asia, and Africa. He has been serving as a reviewer of various national/international journals. He is an active member of AERA and BERA and has presented his papers at various international conferences. He is also the founder of the International Conference on Teaching and Learning (ICOTAL) regularly held at the Islamia University of Bahawalpur, Pakistan. He is a member of the doctoral students selection committee at the Department of Education at his university.

**Yulia Irnidayanti** obtained her first degree in Biology Education and a PhD in Biology. She is currently a senior lecturer and researcher with the Biology and Biology Education Department, Universitas Negeri Jakarta (State University of Jakarta), Indonesia. Since 2001, she has been working together with the Teacher Education Department of the University of Groningen, the Netherlands, on the project about teaching quality and student academic motivation from an international perspective (ICALT3/Differentiation project, Principal Investigator Indonesia). She is interested in helping teachers to improve their teaching quality and student differences in their learning needs, motivation, and learning style.

**Okhwa Lee** is a professor at the Department of Education, Chungbuk National University, South Korea, and CEO of SmartSchool (Ltd). Okhwa Lee is a specialist in educational technology and a practitioner in pre-service teacher education. She is a pioneer of software education, e-learning, and smart education in Korea. She was a member of the Presidential Educational Reform Committee and the Presidential e-Government. She has collaborated in the European Erasmus mobility programme, in research with Finland and the Netherlands, and in the Korean government ODA (Official Development Assistant) programme for Nigeria, Vietnam, and Ethiopia.

**Thelma de Jager** is Head of the Department of Educational Foundation, Tshwane University of Technology, South Africa. She authored and edited textbooks titled *General Subject Didactics and Creative Arts*, in collaboration with other higher education institutions. She has been the author and reviewer of several articles in accredited and peer-reviewed journals. She is involved in the supervision of postgraduate Master’s and Doctoral students. She was awarded women researcher and lecturer of the year several times. Her areas of expertise include teaching and learning strategies and methods to improve the development of learners/students’ critical thinking skills, science education, differentiated teaching, and the inclusion of multicultural students in education. She is also involved in international projects with the British Council (inclusive education), Finland (science teaching), and the Netherlands (comparing quality education with other countries).

**Mattheus Coetzee** is Head of the Department Multimedia Design and Development, Tshwane University of Technology, South Africa. His research interests include teacher education, retention and success rates in engineering studies, as well as statistics of student success. His fields of expertise include academic leadership in higher education, implementation of online teaching and learning technologies, and the development of multimedia for teaching and learning. He was a project leader on an international project to implement eLearning in South African higher education, and currently he leads an Academic Leadership project with two Dutch universities and Southern African institutions.
Nurul Fadhilah is a university lecturer at the Department of Biostatistic and Population, University of Indonesia. She has been actively involved in the international project called ICALT3/Differentiation as an expert observer and as co-investigator for Indonesia. She is currently involved in a research project involving public health big data analysis. She has been involved in professional teacher development for high school teachers in DKI Jakarta. She is experienced in designing and facilitating teacher professional development training, developing syllabi, task designing, and developing differentiated instructions, especially in Cambridge IGCSE and A level Biology.

ORCID
Ridwan Maulana http://orcid.org/0000-0002-0652-7888
Stéfanie André http://orcid.org/0000-0002-5343-0667
Michelle Helms-Lorenz http://orcid.org/0000-0001-9314-6962

References


Opdenakker, M.-C., & Minnaert, A. (2011). Relationship between learning environment characteristics and academic engagement. *Psychological Reports, 109*(1), 259–284. [https://doi.org/10.2466/09.10.11.PR0.109.4.259-284](https://doi.org/10.2466/09.10.11.PR0.109.4.259-284)


### Appendix 1. Example of dimensions and the corresponding items of observation instrument

<table>
<thead>
<tr>
<th>Dimension</th>
<th>High inference indicator</th>
<th>Low inference indicator</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>The teacher…</td>
<td>Rating</td>
<td>The teacher…</td>
</tr>
<tr>
<td>Safe and stimulating learning environment</td>
<td>shows respect for the pupils in behaviour and language use</td>
<td>1 2 3 4</td>
<td>Allows pupils to finish speaking</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Listens to what pupils have to say</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Makes no role-confirming remarks</td>
</tr>
<tr>
<td></td>
<td>Ensures a relaxed atmosphere</td>
<td>1 2 3 4</td>
<td>Addresses the children in a positive manner</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Reacts with humour and stimulates humour</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Demonstrates warmth and empathy toward all pupils</td>
</tr>
<tr>
<td>Efficient classroom management</td>
<td>ensures the lesson proceeds in an orderly manner</td>
<td>1 2 3 4</td>
<td>Intervenes timely and appropriately in case of disorder</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Safeguards the agreed rules and codes of conduct</td>
</tr>
<tr>
<td></td>
<td>uses the time for learning efficiently</td>
<td>1 2 3 4</td>
<td>Starts the lesson on time</td>
</tr>
<tr>
<td>Clear instruction</td>
<td>presents and explains the subject material in a clear manner</td>
<td>1 2 3 4</td>
<td>Does not keep pupils waiting</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Gives staged instructions</td>
</tr>
<tr>
<td></td>
<td>gives clear explanation of how to use didactic aids and how to carry out assignments</td>
<td>1 2 3 4</td>
<td>Poses questions which pupils can understand</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Explains how lesson aims and assignments relate to each other</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Explains clearly which materials and sources can be used</td>
</tr>
<tr>
<td>Activating learning</td>
<td>stimulates pupils to think about solutions</td>
<td>1 2 3 4</td>
<td>Shows pupils the path they can take towards a solution</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Shows learners how to consult sources and reference works</td>
</tr>
<tr>
<td></td>
<td>gives interactive instructions</td>
<td>1 2 3 4</td>
<td>Promotes the interaction between pupils</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Promotes the interaction between teacher and pupils</td>
</tr>
<tr>
<td>Adaptation of teaching</td>
<td>offers weaker learners extra study and instruction time</td>
<td>1 2 3 4</td>
<td>Gives weaker learner extra study time</td>
</tr>
<tr>
<td></td>
<td>adjusts instructions to relevant inter-learner differences</td>
<td>1 2 3 4</td>
<td>Gives weaker learners extra exercises/practices</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Gives additional instructions to small groups or individual pupils</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Does not simply focus on the average learner</td>
</tr>
<tr>
<td>Teaching learning strategy</td>
<td>teaches pupils how to simplify complex problems</td>
<td>1 2 3 4</td>
<td>Teaches pupils how to break down complex problems into simpler ones</td>
</tr>
<tr>
<td></td>
<td>teaches pupils to check solutions</td>
<td>1 2 3 4</td>
<td>Teaches pupils to order complex problems</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Teaches pupils how to estimate outcomes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Teaches pupils how to predict outcomes</td>
</tr>
</tbody>
</table>