Measuring teaching quality and student engagement in South Korea and The Netherlands

van de Grift, Wim J.C.; Chun, Seyeoung; Maulana, Ridwan; Lee, Okhwa; Helms-Lorenz, Michelle

Published in:
School Effectiveness and School Improvement

DOI:
10.1080/09243453.2016.1263215

Publication date:
2017

Routledge
Taylor & Francis Group

School Effectiveness and School Improvement
An International Journal of Research, Policy and Practice

ISSN: 0924-3453 (Print) 1744-5124 (Online) Journal homepage: http://www.tandfonline.com/loi/nses20

Measuring teaching quality and student engagement in South Korea and The Netherlands

Wim J.C.M. van de Grift, Seyeoung Chun, Ridwan Maulana, Okhwa Lee & Michelle Helms-Lorenz

To cite this article: Wim J.C.M. van de Grift, Seyeoung Chun, Ridwan Maulana, Okhwa Lee & Michelle Helms-Lorenz (2016): Measuring teaching quality and student engagement in South Korea and The Netherlands, School Effectiveness and School Improvement, DOI: 10.1080/09243453.2016.1263215

To link to this article: http://dx.doi.org/10.1080/09243453.2016.1263215

Published online: 02 Dec 2016.

SCHOOL EFFECTIVENESS AND SCHOOL IMPROVEMENT, 2016
http://dx.doi.org/10.1080/09243453.2016.1263215

Routledge
Taylor & Francis Group

Measuring teaching quality and student engagement in South Korea and The Netherlands

Wim J.C.M. van de Grift[a], Seyeoung Chun[b], Ridwan Maulana[a], Okhwa Lee[c] and Michelle Helms-Lorenz[a]

[a]Department of Teacher Education, University of Groningen, Groningen, The Netherlands; [b]Department of Education, Chungnam National University, Daejeon, South Korea; [c]Department of Education, Cheongju, Chungbuk National University, South Korea

ABSTRACT

Six observation scales for measuring the skills of teachers and 1 scale for measuring student engagement, assessed in South Korea and The Netherlands, are sufficiently reliable and offer sufficient predictive value for student engagement. A multigroup confirmatory factor analysis shows that the factor loadings and intercepts of the scales are the same, within acceptable boundaries, in both countries. Therefore, we can compare the average scores of teachers in both countries in a reliable and valid way. The 289 Dutch teachers score significantly better on "creating a safe and stimulating learning climate" and "intensive and activating teaching" and almost significantly on "efficient classroom management". We find no significant differences in "clear and structured instruction". The 375 South Korean teachers perform significantly better than the Dutch teachers on "teaching learning strategies" and almost significantly on "differentiating instruction". Furthermore, we find better student engagement in South Korea.

ARTICLE HISTORY
Received 23 May 2016
Accepted 17 November 2016

KEYWORDS
Teacher effectiveness; observation; international comparison; student engagement

Introduction

According to a study of the Programme for International Student Assessment (PISA) of the Organisation for Economic Co-operation and Development (OECD), published in 2012, the average scores earned by Dutch 15-year-old students for reading, mathematics, and science are, respectively, 25%, 31%, and 16% of a standard deviation lower than the average scores of South Korean 15-year-olds. Cohen (1967) evaluates percentages of a standard deviation, or effect sizes (Cohen's δ) between –.19 and .19 as negligible, effect sizes between .20 and .49 as small, effect sizes between .50 and .79 as medium, and effect sizes between .80 and 1.29 as large. From 1.30 on, an effect size is considered very large. Based on meta-analysis, Lipsey (1990) proposes a more empirically detailed set of guidelines for effectiveness research in the behavioural sciences. He evaluates effect sizes (Cohen's δ) between .15 and .44 as small, effect sizes between .45 and .89 as medium, and effect sizes equal or higher than .90 as large. We may conclude that the differences between the scores of the Dutch and the South Korean 15-year-old students for reading, mathematics, and science are small, but nevertheless interesting.

Explanations of these differences in effect sizes lead down the difficult road of comparative international studies. Differences in student achievement might be explained by variations in a vast number of variables, including students' engagement and their individual and social characteristics,

CONTACT Wim J.C.M. van de Grift W.J.C.M.van.de.Grift@rug.nl

distinct cultures and schooling structures, and differences in the quality of teaching. In South Korea, but not in The Netherlands, a growing number of private tutors and out-of-school learning supplements have become a hot topic in the current country's educational debate. Hence, examining student achievement in South Korea should take these factors into account. Additionally, the lack of evidence concerning measurement invariance of the construct being compared in past studies could also be a potential explanation of the differences.

Comparing teaching quality between The Netherlands and South Korea is an interesting endeavour. Both countries are consistently ranked amongst the top 10 with regard to student performance in popular comparative studies such as PISA, TIMSS (Trends in International Mathematics and Science Study), and PIRLS (Progress in International Reading Literacy Study). Compared to The Netherlands, however, South Korea has a relatively higher ranking. Potential differences in the quality of teaching between the two countries can give way to future research revealing the common teaching practice in each specific country.

Nevertheless, this comparison is valid only if there is sufficient evidence regarding the comparability of the construct being compared in the two different contexts (i.e., measurement invariance).

In this study, we concentrate on identifying a method of measuring the quality of teaching in a way that is reliable, valid, and (measurement-)invariant across both countries. Still, we acknowledge that the quality of teaching is just one of the possible explanations for differences in student achievement.

Theoretical and empirical background

International comparisons on the quality of teaching

The TIMSS videotape study of mathematics lessons in different countries reported by Stigler, Gonzales, Kawanaka, Knoll, and Serrano (1999) revealed that students in East Asian (Japan and Hong Kong) countries did not talk a lot in the classroom, and that the students were exposed to more instructional content in comparison to students in other countries. The mathematics problems they worked on were set up mainly using mathematical language. The findings show that high-quality teaching and learning can take place even in a teacher-directed classroom. It is argued that these East Asian classroom practices are deeply rooted in the wider cultural values, with certain practices being embedded in the societal fabric. The authors mention that a simple transplant of those practices from high-achieving countries to low-achieving countries would not work without transplanting the culture as well.

With the help of national school inspectors, Van de Grift (2007) observed the quality of 854 mathematics lessons of 9-year-old pupils in England, Flanders (Belgium), Lower Saxony (Germany), and The Netherlands. He found that the country level could explain only a small, not significant, percentage of teaching quality differences between teachers in primary education. Furthermore, he concluded that the five aspects of quality of teaching ("efficient classroom management", "safe and stimulating learning climate", "clear instruction", "adaptation of teaching", and "teaching-learning strategies") are positively and significantly correlated with pupil involvement, attitude, behaviour, and attainment.

In 2014, Van de Grift reported about the teaching quality observed in large representative samples in primary education from Flanders (Belgium), Lower Saxony (Germany), the Slovak Republic, and The Netherlands. Measures of "creating a safe and stimulating climate", "clear and activating instruction", and "teaching learning strategies" were reliable and fully or at least partially scalar equivalent across these countries. Flemish teachers score higher on creating a safe and stimulating learning climate than teachers in Lower Saxony, the Slovak Republic, or The Netherlands. With regard to the provision of clear and activating instruction, no significant differences arose in average scores across the four countries. Dutch teachers scored significantly

higher on teaching learning strategies compared to teachers in Flanders and Lower Saxony but did not differ significantly from teachers in the Slovak Republic. Flemish and Slovak teachers got higher average scores on teaching learning strategies than teachers in Lower Saxony.

In a small-scale cross-cultural study of gifted students in Singapore and the United States, differences in instructional practices were found between teachers of the two countries (Van Tassel-Baska et al., 2008). The observers rated teachers' curriculum planning and delivery, accommodation for individual differences, problem-solving strategies, critical-thinking strategies, creative-thinking strategies, and research strategies. The study showed that Singapore teachers demonstrated a higher level of effectiveness than American teachers in both general teacher behaviours and differentiation strategies. The level of instructional effectiveness appeared to be positively related to the number of years of teaching experience and training in differentiation practices for the gifted.

In the Teaching And Learning International Study (TALIS; OECD, 2014), teachers of lower secondary education in 24 countries had to fill in questionnaires about self-efficacy in classroom management, instruction, and student engagement. Teachers from The Netherlands showed, on average, more self-efficacy in classroom management, instruction, and student engagement than the teachers in South Korea. The TALIS study is based on self-reports on teaching. Self-reports seem to be vulnerable for social desirability and response style bias.

In the study reported here, we concentrate on observing teaching skills in secondary education. In order to avoid social desirability or response style bias, the study is not based on questionnaires filled in by teachers but on observations made by specially trained external observers.

Teachers' influence on student achievement

Coleman et al. (1966) concluded that most differences in students' achievement can be explained by the students' own intelligence and socioeconomic background; the teacher does not seem to matter in their study. These conclusions have inspired researchers from all over the world to conduct continuing research into the effectiveness of teaching. The earliest contributions to this research stream mostly involved small-scale studies of primary and secondary education, as well as outlier studies into high- and low-achieving schools. These explorations were followed by larger scale surveys with increasing sophistication: Investigations based on single measures of student achievement were followed by studies in which learning gains and added value served as dependent variables. The use of teacher questionnaires was gradually replaced by event sampling and time sampling instruments to observe teachers' behaviours. Simple correlation studies gave way to more complex, multilevel regression analyses. Replacing large-scale surveys, field experiments were conducted by training teachers in an experimental group in some promising behaviours, and then student learning gains were compared in the experimental versus control groups.

The results of these research efforts made clear that about 15% to 25% of the differences in students' achievement might be explained by the work of teachers (Aaronson, Barrow, & Sander, 2007; Bosker & Witziers, 1996; Brandsma & Knuver, 1989; Houtveen & Van de Grift, 2007a, 2007b; Houtveen, Van de Grift, & Brokamp, 2014; Houtveen, Van de Grift, & Creemers, 2004; Rockoff, 2004; Roeleveld, 2003; Wijnstra, Ouwens, & Béguin, 2003). Students of teachers whose teaching quality is a full standard deviation higher, achieve 10% to 25% more learning gains (Aaronson et al., 2007; Bosker & Witziers, 1996; Brandsma & Knuver, 1989; Hanushek & Rivkin, 2010; Houtveen & Van de Grift, 2007a, 2007b; Kane & Staiger, 2008; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004; Roeleveld, 2003; Wijnstra et al., 2003). Thus, it is not only caused by the students' intelligence and socioeconomic background; the teacher matters in determining learning. The societal impact of the learning gains obtained through skilled teachers also has been estimated by educational economists, who conclude that students of these better skilled teachers earn annually, on average, $20,000 more, later on (Hanushek, 2011; Hanushek & Rivkin, 2010).

In turn, the key question becomes what effective teachers do in the classroom to provide these benefits. As research reviews make clear, several teaching behaviours encourage more student achievement and more learning gains, including setting targets, offering sufficient learning and instruction time, monitoring students' achievements, creating special measures for struggling learners, establishing a safe and stimulating educational climate, organizing efficient classroom management, giving clear and structured instruction, organizing intensive and activating teaching, differentiating instruction, and teaching learning strategies (Cotton, 1995; Creemers, 1991, 1994; Ellis & Worthington, 1994; Levine & Lezotte, 1990, 1995; Muijs & Reynolds, 2011; Purkey & Smith, 1983; Sammons, Hillman, & Mortimore, 1995; Scheerens, 1989, 1992, 2008; Van de Grift, 1985, 1990; Walberg & Haertel, 1992; Wright, Horn, & Sanders, 1997). These results have been popularized in several books (Hattie, 2009, 2012; Marzano, 2003).

However, it is not possible to observe all of the mentioned teaching behaviours during the lesson. In the context of (natural setting) observations, the following six behaviours are observable: establishing a safe and stimulating educational climate, organizing efficient classroom management, giving clear and structured instruction, organizing intensive and activating teaching, differentiating instruction, and teaching learning strategies (Maulana, Helms-Lorenz, & Van de Grift, 2015; Van de Grift, 2007). Similarly, three domains of instructional quality (cf. teaching behaviour) are often mentioned in comparison studies and educational psychology literature. These domains include supportive classroom climate, classroom management, and cognitive activation (Baumert et al., 2010; Klieme, Pauli, & Reusser, 2009). Supportive classroom climate and classroom management are mentioned distinctively in both Van de Grift's (2007) as well as Klieme, Pauli, and Reusser's (2009) conceptualizations of teaching. The domain of cognitive activation mentioned by Klieme et al. (2009) coincides with the remaining domains of Van de Grift (2007).

Research questions

Comparing teaching quality across two countries presupposes that the measures in both countries are reliable and have sufficient predictive validity. It is also important to minimize measurement variance between the samples of the countries; that is, the items in the measurement scales need to have the same meaning in both countries. Otherwise, any comparison is meaningless. The research questions for this study thus are:

- Are the measures of teaching quality and student engagement in South Korea and The Netherlands reliable in both countries?
- Are the instrument items psychometrically equal in both countries?
- Are the measures of teaching quality and student engagement valid in both countries?

Method

Samples

For the South Korean data, the teaching skills of a sample of 375 teachers working in 26 secondary schools in Daejeon (241), Chungnam (61), and Chungbuk (74) were observed. These teachers teach 25 different subjects. The Dutch data were gathered by specially trained Dutch school inspectors, across a random sample of 76 departments of secondary schools all over the country. In each school, about 4 teachers were observed. The 289 Dutch teachers teach 21 different subjects. Table 1 presents some background variable scores for both countries, as well as the significance level of difference between the sample characteristics across the two countries. As can be seen from Table 1, the sample characteristics do not differ across the countries for gender, but with a *t*

Table 1. Sample characteristics.

	South Korea n = 375	Netherlands n = 289	Significance (t test)	Effect size (Cohen's δ)
Percentage of male teachers	49	50	.802	.020
Years of experience	11.31	16.53	.000	.510
Class size	29.11	21.70	.000	1.149

test significant differences are found for years of experience and class size. Dutch teachers have more years of experience (Cohen's δ = .51) and much smaller classrooms (Cohen's δ = 1.15).

The sample sizes of Dutch and South Korean teachers are large enough (n is 375 and 289, respectively) to find significant (< .05 with a power of .80) differences with effect sizes (Cohen's δ) of about .20 and more.

Instruments

We obtained, from some of the educational effectiveness studies we reviewed previously, 32 high-inferential observable teaching activities and 120 low-inferential observable teaching activities. These activities are arranged into six categories: a safe and stimulating educational climate, efficient classroom management, clear and structured instruction, intensive and activating teaching, teaching learning strategies, and differentiating instruction. The same items had been used previously to construct the International Comparative Analysis of Learning and Teaching (ICALT) instrument for observing teachers in primary education (Van de Grift, 2007, 2014; Van de Grift & Lam, 1998; Van de Grift, Van der Wal, & Torenbeek, 2011), beginning teachers in secondary education (Van de Grift, Helms-Lorenz, & Maulana, 2014), and experienced teachers in secondary education (Van der Lans, Van de Grift, Van Veen, & Fokkens-Bruinsma, 2016). Observers rate the items on a 4-point scale, and scores for each scale reflect the average item score (1–2 = insufficient; 2–3 = sufficient, and 3–4 = good).

Van de Grift developed with the Inspectorate of Education (Inspectie van het Onderwijs, 2009) in The Netherlands a three-item scale for observing student academic engagement in a lesson. This Likert-style scale measures students' academic engagement, with an emphasis on psychological and behavioural engagement, using items such as, "Students are fully engaged in learning" and "Students show that they are interested in learning". Observers rated the items on a 4-point response scale, ranging from 1 (*Completely not true*) to 4 (*Completely true*).

Reliability

The reliability, construct validity, and intercultural equivalence of the factor structure of the ICALT scales and the student engagement scale were tested previously, for studies conducted in The Netherlands, Germany, Flanders, England, Scotland, and Slovak Republic (Van de Grift, 2007, 2014; Van de Grift & Lam, 1998; Van de Grift et al., 2011).

Predictive validity

With the first ICALT version, Van de Grift and Lam (1998) gathered observation data from 884 teachers of 227 primary schools and related these observations to the school-level corrected output. In a multilevel regression analysis, they found a significant beta (.16) for teaching skill. The combined effect of the school surroundings and teaching skill explained about 36% of the between-school variance (Van de Grift & Lam, 1998).

In secondary education, with samples of teachers providing instruction in more than 20 different subjects, it becomes virtually impossible to relate teaching skill to school-level student achievement outputs. As an alternative, some studies computed the correlation of the ICALT instrument

with Van de Grift's students' academic engagement scale (Van de Grift, 2007). Pre-service teacher behaviour uniquely explained about 27% of the variance in class academic engagement (Van de Grift et al., 2014).

Translation

To measure teaching quality in South Korea, we constructed the Korean version of the ICALT observation instrument following the guidelines of the International Test Commission (Hambleton, 1994). The process included the translation and back-translation of the instrument. The English version of the ICALT instrument (Van de Grift et al., 2014) was used as the source language for translation. This process involved two highly knowledgeable researchers with respect to the instrument and the theoretical framework underlying the instrument and two Korean professors proficient in both English and Korean languages. Upon completion of the translation and back-translation procedure, minor discrepancies were discussed thoroughly and resolved subsequently.

Training of observers

For the Dutch secondary education setting, the observations were conducted by 13 inspectors of the Inspectorate of Education in The Netherlands. These school inspectors were trained by using videotaped lessons. They reached a Cohen's κ of .76 when observing the same teacher, representing good mutual consensus.

The South Korean observers were trained over the course of a full day by two English-speaking Dutch trainers. The training involved explanations of the observation instruments, two videotaped lessons, and a discussion about how to evaluate teaching practices using the associated scoring rules. After discussion, the South Korean observers reached a Cohen's κ of .60 (moderate/good mutual consensus).

Statistical analyses

Reliability

Table 2 presents some classical reliability coefficients for the six ICALT scales and the scale for measuring student engagement. In both South Korea and The Netherlands, all ICALT scales and the scale for measuring student engagement thus are sufficiently homogeneous (Cronbach's α >.70).

Comparing average sum scores is very common in educational research. For a fair comparison, the factor loadings and the intercepts of the indicators have to be equal in both South Korea and The Netherlands. A fair comparison asks for measurement invariance between both countries.

Table 2. Reliability of scales for measuring teaching skills and student engagement in South Korea and The Netherlands.

	Basic teaching skills			Advanced teaching skills				
	Safe and stimulating learning climate	Efficient classroom management	Clear and structured Instruction	Intensive and activating teaching		Teaching learning strategies	Differen-tiating instruction	Student engagement
Cronbach's α	(4)[1]	(4)	(7)	(7)	(6)[2]	(6)	(4)	(3)
South Korea	.82	.80	.87	.83	.83	.86	.84	.85
Netherlands	.90	.72	.86	.85	.84	.82	.73	.87

[1]Number of items in the scale. [2]Leaving out Item 22: "explains the lesson objectives at the start of the lesson".

Measurement invariance between South Korea and The Netherlands

Observers and teachers in different cultures and countries should ascribe the same meaning to the same scale items. If measurement variance arises between cultures or countries, the comparison of the results becomes useless. Measurement invariance refers to "whether or not, under different conditions of observing and studying a phenomenon, measurement operations yield measures of the same attribute" (Horn & McArdle, 1992, p. 117). Classical tests consider several levels of measurement equivalence. For example, configural equivalence means that items loading strongly on the latent factor in one group also load high on that factor in other groups. Metric equivalence presupposes that the factor loadings of the items are equal for all groups, which is important for comparing regression coefficients across groups. The most appropriate level for comparing the average scores of groups, scalar equivalence, is also the most stringent. It means that both the factor loadings and the intercepts of the indicators are equal in all groups.

The usual χ^2-based test for comparing parameters is substantially affected by sample size (Marsh, Balla, & McDonald, 1988). Because we have large samples of observations, we use the comparative fit index (CFI) and the Tucker-Lewis index (TLI). Both indices are less vulnerable to sample size. Furthermore, we consider the root mean square error of approximation (RMSEA) to assess model fit. The norms for acceptable fit are CFI and TLI > .90 and RMSEA < .08 (Chen, Curran, Bollen, Kirby, & Paxton, 2008; Hu & Bentler, 1999; Kline, 2005; Marsh, Hau, & Wen, 2004; Tucker & Lewis, 1973).

We used the program Mplus 7.4 for the tests for configural, metric, and scalar equivalence for the six ICALT scales and the scale for measuring student engagement in both countries. The results are shown in Table 3.

The multigroup confirmatory factor analysis (MGCFA) for the six ICALT scales and the scale for student engagement with South Korean and Dutch teachers resulted for configural and metric equivalence in a CFI > .90, a TLI > .90, and a RMSEA < .08, which indicates acceptable fit. Scalar equivalence was CFI > .90 and TLI > .90, but the RMSEA was exactly .08. From the study of Van de Grift et al. (2014), with the same items but another set of data, we remember that Item 22, "explains the lesson objectives at the start of the lesson", violates several model assumptions. Therefore, Item 22 was left out. Without Item 22, the multigroup confirmatory factor analysis for the six ICALT scales and the scale for student engagement with South Korean and Dutch teachers resulted for configural, metric, and scalar equivalence in a CFI and TLI > .90 and a RMSEA < .08, all of which indicate acceptable fit. Therefore, the six ICALT scales and the scale for student engagement support valid comparisons of the average scores across the two countries.

The decision for a specific model is based on the differences in the goodness-of-fit indices (GFI). Cheung and Rensvold (2002) and Chen (2007) have conducted Monte Carlo studies in order to examine how goodness-of-fit indices change when between-group constraints are added to a measurement model. More specifically, they examined changes in the goodness-of-fit indices when invariance constraints (configural, metric, scalar) were added. These authors proposed critical values of these so-called ΔGFI's that indicate measurement invariance. Norms for decrease in model fit indices are: ΔCFI < .01, ΔTLI < .01, ΔRMSEA < .015 (Chen, 2007; Cheung & Rensvold, 2002).

Table 4 shows that the results with ΔRMSEA are all below the norm of .015. The results with: ΔCFI and ΔTLI are almost below .01.

Table 3. Multigroup confirmatory factor analysis (MGCFA) on 7 factors in South Korea and The Netherlands.

	35 items 7 factors			34 items 7 factors (Item 23 removed)		
Model fit Norm	CFI > .90	TLI > .90	RMSEA < .08	CFI > .90	TLI > .90	RMSEA < .08
Configural	.934	.927	.072	.936	.929	.072
Metric	.930	.925	.073	.932	.927	.073
Scalar	.911	.909	.080	.916	.915	.079

Table 4. Nested model comparisons of multigroup confirmatory factor analysis.

Models comparison	ΔCFI	ΔTLI	ΔRMSEA	Decision
35 items 7 factors				
Metric versus configural	−.004	−.002	.001	Accept
Scalar versus metric	−.019	−.016	.007	Accept
Scalar versus configural	−.023	−.017	.008	Accept
34 items 7 factors				
Metric versus configural	−.004	−.004	.001	Accept
Scalar versus metric	−.016	−.012	.006	Accept
Scalar versus configural	−.020	−.014	.007	Accept

Norm for decrease in model fit indices: ΔCFI < 0.01, ΔTLI < 0.01, ΔRMSEA < 0.015 (Chen, 2007; Cheung & Rensvold, 2002).

Predictive validity

In international comparisons, with samples of teachers who provide instruction in a lot of different subjects, it becomes virtually impossible to relate teaching skills to school-level student achievement outputs. As an alternative, we studied the relationships of the six ICALT scales with students' academic engagement. We used Van de Grifts three-item scale for observing student academic engagement in a lesson (Inspectie van het Onderwijs, 2009). With Mplus 7.4, we computed the γ coefficients between the six latent teaching skills and student engagement in both South Korea and The Netherlands. Table 5 presents the results.

In both countries, the γ coefficients between the latent teaching skills and student engagement are high (γ ≥ .50) and significant (p > .001). Teachers with high teaching skills have more engaged students in their classrooms. This is an important indication of the predictive validity of the six teaching skills measured with the ICALT instrument.

Conclusions

Six ICALT scales for measuring teaching skills, assessed in South Korea and The Netherlands, are sufficiently reliable and offer sufficient predictive value for student engagement. A multigroup confirmatory factor analysis shows that the factor loadings and intercepts of the six ICALT scales are the same, within acceptable boundaries, in both countries. Therefore, we can compare the average scores of teachers in both countries in a reliable and valid way.

Descriptive statistics

Table 6 presents some descriptive statistics of the sum scores of the ICALT scales and the scale for measuring student engagement.

Dutch teachers significantly (< .001) outperform South Korean teachers with regard to creating a "safe and stimulating learning climate", with a small effect size (Cohen's δ) of .37. The difference in "efficient classroom management" (Cohen's δ = .18) in favour of the Dutch teachers is with a significance level of .052 just not significant. The negligible difference (Cohen's δ = .10) in "clear

Table 5. γ coefficients between latent teaching skills and student engagement in South Korea and The Netherlands.

	Basic teaching skills		Advanced teaching skills			
	Safe and stimulating learning climate (4)	Efficient classroom management (4)	Clear and structured Instruction (7)	Intensive and activating teaching (6)	Teaching learning strategies (6)	Differentiating instruction (4)
γ with student engagement in:						
South Korea	.76	.80	.86	.87	.73	.72
Netherlands	.77	.79	.77	.74	.50	.51

Table 6. Comparison of sum scores on ICALT scales in South Korea and The Netherlands.

	Basic teaching skills						Advanced teaching skills							
	Safe and stimulating lear-ning climate (4)		Efficient classroom management (4)		Clear and structured instruction (7)		Intensive and activating teaching (6)		Teaching learning strategies (6)		Differentiating instruction (4)		Student engagement (3)	
	Avg	*SD*	Avg	*SD*	Avg	*SD*	Avg	*SD*	Avg	*SD*	Avg	*SD*	Avg	*SD*
South Korea	3.04	.67	3.05	.69	2.91	.64	2.69	.67	2.57	.73	2.36	.83	3.09	.70
Netherlands	3.27	.55	3.16	.51	2.97	.52	2.82	.56	2.45	.53	2.26	.46	2.91	.62
Effect size	.37		.18		.10		.21		−.18		−.14		−.27	
Significant	.000		.052		.155		.009		.024		.054		.001	

and structured instruction" between both countries is not significant. Dutch teachers significantly (< .001) outperform South Korean teachers with regard to organizing "intensive and activating teaching" (Cohen's δ = .21). South Korean teachers outperform Dutch teachers significantly (< .05) when it comes to "teaching learning strategies" (Cohen's δ = .18) The difference in "differentiating instruction" (effect size = .14) in favour of the Korean teachers is with a significance level of .054 just not significant. The South Korean students are significantly (.001) more (effect size = .27) engaged in the lessons than the Dutch students.

Discussion

The main focus of this study was to examine the equality of psychometric qualities of observation instruments in two vastly different contexts. This was established, paving the way to a first cautious comparison. Even though our study differs in terms of sample size and measurement instruments, our findings are in line with the scant evidence found in international comparative studies (Stigler et al., 1999; Van Tassel-Baska et al., 2008), showing higher levels of higher order teaching skills of East Asian teachers.

Sample size

In this study, we found only small differences (Cohen's δ < .40) in the teaching skills of Dutch and South Korean teachers. Some differences between both countries were just not significant at the .05 level, while the effect sizes of these differences are still interesting. For example, for "efficient classroom management" and "differentiating instruction", we found just not significant effect sizes of Cohen's δ .18 and .14. For finding effect sizes of > .15 significant at the .05 level with a power of .80, two samples of each more than 550 teachers are needed. We should keep this in mind for future studies on international comparisons.

Inter-rater reliability

Using observations in international comparative studies is vulnerable to cultural differences. Even though indigenous observers collected the data, future studies should investigate the inter-rater reliability using a cross-over design of observers from both countries rating teachers from their own as well as the foreign country. This should minimize cultural observation biases.

Explaining differences between countries

This study was inspired by the differences in the achievements of South Korean and Dutch students, as indicated in international comparative studies by the OECD. The behaviour observed with the ICALT instrument has been established in the literature as being effective, meaning that it

contributes to student achievement. We acknowledge that not all effective teacher behaviour is captured with this instrument, but for our research questions this instrument seems to be sufficient. We are also aware of the fact that only 15% to 25% of the differences in student achievement may be ascribed to the work of teachers. Other explanatory factors that need attention in follow-up studies are:

- demographical homogeneity (Berry & Ward, 2016);
- cultural settings (King & Bernardo, 2016);
- motivation and attitude of students (Lee, 2013; Maulana, Helms-Lorenz, & Van de Grift, 2016; Shin & Han, 2013);
- private tutoring (Kim & Lee, 2010);
- amount of illiteracy (Lee, 2013).

Nevertheless, this study offers some evidence that South Korean students have access to teachers with more advanced teaching skills with regard to teaching students how they should learn and better differentiating their instruction to meet the distinct needs of their students. These advanced teaching skills have great potential to influence the learning gains of both struggling and excellent learners. This might also contribute, *amongst other factors*, to the higher level of student engagement evident from our findings in the South Korean sample.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Wim J.C.M. van de Grift is full professor in Educational Sciences at the University of Groningen. He is a member of the Dutch research school in the domain of education (ICO) and scientific advisor of the Inspectorate of Education in The Netherlands. Van de Grift's research programme is aimed at the development and testing of theories on the professional development of teachers in different countries. This research programme focuses on the following questions: How do teaching skills develop during the teaching career? Which factors influence the development of teaching skills? What is the influence of teachers' teaching skills on students' academic engagement and achievements?

Seyeoung Chun, PhD, has teached at the Department of Education of the Chungnam National University of South Korea since 1997. He received his education and PhD from Seoul National University and has been actively doing research in educational policy and has had several key positions, such as Secretary of Education to the President and CEO of the Korean Educational Research and Information Service (KERIS). Professor Chun is the founder and current president of Smart Education Society with more than 3,000 members.

Ridwan Maulana is assistant professor at the Department of Teacher Education, University of Groningen, The Netherlands. His major research interests involve teaching and teacher education, factors influencing teaching quality, statistics, and methods associated with the measurement of teaching, longitudinal research, cross-country comparisons, and effects of teaching behaviour on students' motivation and engagement. He has been involved in various teacher professional development projects including the Dutch induction programme (LONIE) and school-university-based partnership (OIDS). He is currently a project leader of an international project on teaching quality (ICALT3/ Differentiation) involving countries from Europe, Asia, and Africa.

Okhwa Lee, PhD, is 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.

Michelle Helms-Lorenz is associate professor at the Department of Teacher Education, University of Groningen, The Netherlands. Her research interest in the cultural specifity versus universality (of behaviour and psychological

processes) was fed by the cultural diversity in South Africa, where she was born and raised. Michelle's second passion is education, the bumpy road toward 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 (PD) projects including the Dutch Induction programme (LONIE), school-university-based partnership (OIDS), and the in-service PD programme (OOD).

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*, 95–135.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., … Tsai, Y.-M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, *47*, 133–180. doi:10.3102/0002831209345157

Berry, J. W., & Ward, C. (2016). Multiculturalism. In D. L. Sam & J. W. Berry (Eds.), *The Cambridge handbook of acculturation psychology* (2nd ed., pp. 441–463). Cambridge, UK: Cambridge University Press.

Bosker, R. J., & Witziers, B. (1996, April). *The magnitude of school effects, or: Does it really matter which school a student attends?* Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.

Brandsma, H. P., & Knuver, J. W. M. (1989). Effects of school and classroom characteristics on pupil progress in language and arithmetic. *International Journal of Educational Research*, *13*, 777–788.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*, 464–504.

Chen, F., Curran, P .J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research*, *36*, 462–494.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 233–255.

Cohen, J. (1967). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity*. Washington, DC: US Department of Health, Education & Welfare/Office of Education.

Cotton, K. (1995). *Effective schooling practices: A research synthesis 1995 update*. Portland, OR: Northwest Regional Educational Laboratory.

Creemers, B. P. M. (1991). *Effectieve instructie* [Effective Instruction]. Den Haag, The Netherlands: SVO.

Creemers, B. P. M. (1994). *The effective classroom*. London, UK: Cassell.

Ellis, E. S., & Worthington, L. A. (1994). *Research synthesis on effective teaching principles and the design of quality tools for educators* (Technical Report No. 5). Eugene, OR: University of Oregon, National Center to Improve the Tools of Educators.

Hambleton, R. K. (1994). The rise and fall of criterion-referenced measurement? *Educational Measurement: Issues and Practice*, *13*(4), 21–26. doi:10.1111/j.1745-3992.1994.tb00567.x

Hanushek, E. A., (2011). The economic value of higher teacher quality. *Economics of Education Review*, *30*, 466–479.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review: Papers & Proceedings*, *100*(2), 267–271.

Hattie, J. A. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York, NY: Routledge.

Hattie, J. A. (2012). *Visible learning for teachers: Maximizing impact on learning*. New York, NY: Routledge.

Horn, J. L., & McArdle J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*, 117–144.

Houtveen, A. A. M., & Van de Grift, W. J. C. M. (2007a). Effects of metacognitive strategy instruction and instruction time on reading comprehension. *School Effectiveness and School Improvement*, *18*, 173–190.

Houtveen, A. A. M., & Van de Grift, W. J. C. M. (2007b). Reading instruction for struggling learners. *Journal of Education for Students Placed At Risk*, *12*, 405–424.

Houtveen, A. A. M., Van de Grift, W. J. C. M., & Brokamp, S. K. (2014). Fluent reading in special elementary education. *School Effectiveness and School Improvement*, *25*, 555–569.

Houtveen, A. A. M., Van de Grift, W. J. C. M., & Creemers, B. P. M. (2004). Effective school improvement in mathematics. *School Effectiveness and School Improvement*, *15*, 337–376.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55.

Inspectie van het Onderwijs. (2009). *International comparative analysis of learning and teaching in math lessons in several European countries*. Utrecht, The Netherlands: Author.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (NBER Working Paper No. 14607). Retrieved from http://www.nber.org/papers/w14607

Kim, S., & Lee, J.-H. (2010). Private tutoring and demand for education in South Korea. *Economic Development and Cultural Change*, *58*, 259–296.

King, R. B., & Bernardo, A. B. I. (Eds.). (2016). *The psychology of Asian learners: A festschrift in honor of David Watkins*. London, UK: Springer.

Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras study: Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In T. Janik & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 137–160). Münster, Germany: Waxmann.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: The Guilford Press.

Lee, J.-H. (2013). *Positive changes: The education, science & technology policies of Korea*. Seoel, South Korea: Korean Economic Daily & Business Publications.

Levine, D. U., & Lezotte, L.W. (1990). *Unusually effective schools: A review and analysis of research and practice*. Madison, WI: The National Center for Effective Schools Research and Development.

Levine, D. U., & Lezotte, L. W. (1995). Effective schools research. In J. A. Banks & C. A. M. Banks (Eds.), *Handbook of research on multicultural education* (pp. 525–547). New York, NY: Macmillan.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, *103*, 391–410.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers of overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*, 320–341.

Marzano, R. J. (2003). *What works in schools: Translating research into action*. Alexandria, VA: ASCD.

Maulana, R., Helms-Lorenz, M., & Van de Grift, W. (2015). Development and evaluation of a questionnaire measuring pre-service teachers' teaching behaviour: A Rasch modelling approach. *School Effectiveness and School Improvement*, *26*, 169–194.

Maulana, R., Helms-Lorenz, M., & Van de Grift, W. (2016). The role of autonomous motivation for academic engagement of indonesian secondary school students: A multilevel modelling approach. In R. B. King & A. B. I. Bernardo (Eds.), *The psychology of Asian learners: A festschrift in honor of David Watkins* (pp. 237–252). London, UK: Springer.

Muijs, D., & Reynolds, D. (Eds.). (2011). *Effective teaching: Evidence and practice* (3rd ed.). London, UK: Sage.

Organisation for Economic Co-operation and Development. (2012). *PISA 2012 results in focus: What 15-year-olds know and what they can do with what they know*. Paris, France: Author.

Organisation for Economic Co-operation and Development. (2014). *TALIS 2013 results: An international perspective on teaching and learning*. Paris, France: Author.

Purkey, S. C., & Smith, M. S. (1983). Effective schools: A review. *The Elementary School Journal*, *83*, 427–452.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*, 417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, *94*, 247–252.

Roeleveld, J. (2003). *Herkomstkenmerken en begintoets: Secundaire analyses op het PRIMA-cohort onderzoek* [Social background and testing in the early years: Secondary analyses on the PRIMA-cohort study]. Amsterdam, The Netherlands: SCO Kohnstamm Instituut.

Sammons, P. Hillman, J., & Mortimore, P. (1995). *Key characteristics of effective schools: A review of school effectiveness research*. London, UK: Office for Standards in Education.

Scheerens, J. (1989). *Wat maakt scholen effectief? Samenvattingen en analyses van onderzoeksresultaten* [What explains a school's effectivity? Summaries and analyses of research outcomes]. Den Haag, The Netherlands: Instituut voor Onderzoek van het Onderwijs SVO.

Scheerens, J. (1992). *Effective schooling: Research, theory and practice*. London, UK: Cassell.

Scheerens, J. (2008). *Een overzichtsstudie naar school- en instructie-effectiviteit: Samenvattingen en analyses van onderzoeksresultaten* [Review of school and instruction effectivenes: Summaries and analyses of research outcomes]. Enschede, The Netherlands: Universiteit Twente.

Shin, I.-H., & Han, S.-S. (2013). Pulling students out of underachievement. In J. H. Lee (Ed.), *Positive changes: The education, science & technology policies of Korea* (pp. 162–178). Seoul, South Korea: Korean Economic Daily & Business Publications.

Stigler, J. W., Gonzales, P., Kawanaka, T., Knoll, S., & Serrano, A. (1999). *The TIMSS Videotape Classroom Study: Methods and findings from an exploratory research project on eighth-grade mathematics instruction in* Germany, Japan, *and the United States* (NCES 99-074). Washington, DC: U.S. Government Printing Office, 1999.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*, 1–10.

Van de Grift, W. (1985). Onderwijsleerklimaat en leerlingprestaties [Educational climate and student achievement]. *Pedagogische Studiën*, *62*, 401–414.

Van de Grift, W. (1990). Het onderzoek naar effectieve scholen [Studies into the effectiveness of schools]. *Pedagogische Studiën*, *67*, 462–463.

Van de Grift, W. (2007). Quality of teaching in four European countries: A review of the literature and an application of an assessment instrument. *Educational Research*, *49*, 127–152.

Van de Grift, W. J. C. M. (2014). Measuring teaching quality in several European countries. *School Effectiveness and School Improvement*, *25*, 295–311.

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

Van de Grift, W. J. C. M., & Lam, J. F. (1998). Het didactisch handelen in het basisonderwijs [Teaching in primary education]. *Tijdschrift voor Onderwijsresearch*, *23*, 224–241.

Van de Grift, W., Van der Wal, M., & Torenbeek, M. (2011). Ontwikkeling in de pedagogisch didactische vaardigheid van leraren in het basisonderwijs [Development of teaching skills in primary education]. *Pedagogische Studiën*, *88*, 416–432.

Van der Lans, R. M., Van de Grift, W. J. C. M., Van Veen, K., & Fokkens-Bruinsma, M. (2016). Once is not enough: Establishing reliability criteria for feedback and evaluation decisions based on classroom observations. *Studies in Educational Evaluation*. Advance online publication. doi:10.1016/j.stueduc.2016.08.001

Van Tassel-Baska, J., Feng, A., MacFarlane, B., Heng, M. A., Teo, C. T., Wong, M. L., … Khong, C. (2008). A cross cultural study of teachers' instructional practices in Singapore and the United States. *Journal for the Education of the Gifted*, *31*, 338–363.

Walberg, H. J., & Haertel, G. D. (1992). Educational psychology's first century. *Journal of Educational Psychology*, *84*, 6–19.

Wright, S. P., Horn, S. P., & Sanders, W.L. (1997). Teacher and classroom context effects on student achievement: implications for teacher evaluation. *Journal of Personnel Evaluation in Education*, *11*, 57–67.

Wijnstra, J., Ouwens, M., & Béguin, A. (2003). *De toegevoegde waarde van de basisschool* [Added value of schools in elementary education]. Arnhem, The Netherlands: Citogroep.