Fischer, Jessica; Praetorius, Anna-Katharina; Klieme, Eckhard

The impact of linguistic similarity on cross-cultural comparability of students' perceptions of teaching quality

formal und inhaltlich überarbeitete Version der Originalveröffentlichung in:
formally and content revised edition of the original source in:
Educational assessment, evaluation and accountability 31 (2019) 2, S. 201-220

The impact of linguistic similarity on cross-cultural comparability of students' perceptions of teaching quality

Jessica Fischer[1] & Anna-Katharina Praetorius[2] & Eckhard Klieme[1]

Abstract

Valid cross-country comparisons of student learning and pivotal factors contributing to it, such as teaching quality, offer the possibility to learn from outstandingly effective educational systems across the world and to improve learning in classrooms by providing policy relevant information. Yet, it often remains unclear whether the instruments used in international large-scale assessments work similarly across different cultural and linguistic groups, and thus can be used for comparing them. Using PISA 2012 data, we investigated data comparability of three teaching quality dimensions, namely student support, classroom management, and cognitive activation using a newly developed psychometric approach, namely alignment. Focusing on 15 countries, grouped into five linguistic clusters, we secondly assessed the impact of linguistic similarity on data comparability. Main findings include that (1) comparability of teaching quality measures is limited when comparing linguistically diverse countries; (2) the level of comparability varies across dimensions; (3) linguistic similarity considerably enhances the degree of comparability, except across the Chinese-speaking countries. Our study illustrates new and more flexible possibilities to test for data comparability and outlines the importance to consider cultural and linguistic differences when comparing teaching-related measures across groups. We discuss possible sources of lacking data comparability and implications for comparative educational research.

Keywords Data comparability· Teaching quality· Alignment · Linguistic similarity· PISA 2012 · Large-scale assessment

Jessica Fischer: jessica.fischer@dipf.de

[1] DIPF | Leibniz Institute for Research and Information in Education, Frankfurt am Main, Germany

[2] University of Zurich, Zurich, Switzerland

1 Introduction

One of the most important goals of comparative educational research is to explain *why* student achievement varies across countries. By identifying factors that positively influence learning outcomes, policy-relevant information can be gained on how to improve learning in classrooms (van de Vijver and He 2016). Furthermore, impulses are derived on how to learn from outstandingly effective educational systems across the world (Schulz 2003). In the last decade, teaching quality has gained considerable attention in large-scale studies as one of the most important contextual factors (e.g., PISA: Kuger et al. 2017; TALIS 2008 2013: He and Kubacka 2015; TIMSS: Nilsen and Gustafsson 2016). Such contextual factors are often assessed via questionnaires. Yet, comparability of questionnaire measures can be challenged by the diversity of countries participating in large-scale studies (van de Vijver and Leung 1997). Hence, valid comparative inferences require the demonstration of cross-country measurement invariance to ensure that variation lies in the targeted construct rather than being due to non-invariance of measures.

Despite its critical relevance, testing for measurement invariance is often neglected with respect to teaching quality. In the few existing studies, measurement invariance is estimated across the whole sample of participating countries (see, e.g., He and Kubacka 2015). We argue, however, that measurement invariance is heavily dependent on respondents' cultural similarity (see also van de Vijver and Leung 1997). We test our assumption using a purposefully selected sub-sample of countries that participated in PISA 2012. Since traditional methods, and especially multigroup confirmatory factor analysis, have been criticized to be overly strict in large-scale comparisons involving many cultures, we use a more flexible and advanced method, namely alignment.

After introducing the basic dimensions as a model for conceptualizing high-quality teaching (Section 1.1) as well as describing the levels of measurement invariance usually distinguished (Section 1.2), we summarize empirical investigations on invariance of teaching quality measures (Section 1.3). Based on these findings, we derive the research hypotheses for our study (Section 1.4).

1.1 Conceptualizing teaching quality: Three Basic Dimensions

Based on the educational effectiveness paradigm, teaching quality can be defined as instructional aspects influencing students' cognitive and affective learning outcomes (Seidel and Shavelson 2007). Teaching quality can be conceptualized in different ways. One prominent approach is the framework of Three Basic Dimensions (Klieme et al. 2009), comprising the dimensions *student support*, *classroom management*, and *cognitive activation* (for a review see Praetorius et al. 2018b).

Student support refers to instruction characterized by fostering a warm and appreciative teacher-student relationship, providing constructive feedback, individual support, and positively dealing with student errors (e.g., Klieme et al. 2009; Klusmann et al. 2008; Lipowsky et al. 2009). Referring to self-determination theory (Deci and Ryan 1996), students' (intrinsic) motivation to learn (Dietrich et al. 2015), subject-

specific interest (Fauth et al. 2014), and self-concept (Gläser-Zikuda et al. 2017) should be enhanced when students are supported in their learning during instruction.

Classroom management refers to quality learning time. In the sense of Kounin (1970), it does not just relate to the teachers' reaction to disruptions but also to instruction that aims to prevent the occurrence of disruptions in classroom, for instance by effective use of time or clearly defined rules (Praetorius et al. 2018b). Effective management is assumed to influence students' motivational and cognitive learning (e.g., Brophy 2000; Hattie 2009; Kunter et al. 2007; Walberg and Paik 2000).

Cognitive activation summarizes instructional practices promoting students' higher-level thinking and supporting metacognition by using challenging tasks and questions or by activating and exploring students' prior knowledge (Fauth et al. 2014; Klieme et al. 2009; Pinger et al. 2017). Consequently, cognitively activating instruction is assumed to influence cognitive student outcomes (Baumert et al. 2010; Lipowsky et al. 2009) for instance by stimulating students' potential to reconstruct, elaborate, and integrate information (Praetorius et al. 2018b).

Having originally been developed based on German classroom samples, the Three Basic Dimensions are meanwhile prominent in international publications (e.g., Fauth et al. 2014; Fischer et al. 2014; Lipowsky et al. 2009; Nilsen and Gustafsson 2016; Praetorius et al. 2014; Yi and Lee 2017). Yet, it remains unclear whether differences and similarities across countries with respect to the Three Basic Dimensions can be interpreted validly if invariance of measures is not checked carefully prior to comparing the data (Scherer et al. 2016).

1.2 Traditional levels of measurement invariance

If bias is present, score differences from the assessment do not reflect real cross-country differences in the targeted construct (e.g., student support), but are caused by not intended cultural variation affecting survey response. Three types of bias can be distinguished: cross-country differences in (1) construct meaning (construct bias), (2) sampling or respondents' use of the instrument (method bias), and (3) item meaning (item bias). Levels of comparability need to be assessed to check different types of bias and to ensure cross-country comparability of measurements. Traditionally, three hierarchically linked measurement invariance levels can be distinguished (for an overview of bias and equivalence, see van de Vijver and Leung 1997 or He and Kubacka 2015).

Configural invariance indicates that a construct is measured across countries by the same items. When configural invariance is met, the basic structure of a construct can be studied across countries. *Metric invariance* means that not only the same items can be used across countries, but that a construct is also measured by the same metric. Consequently, associations (e.g., correlations) of metric-invariant measures, such as student support and student outcomes, can be compared across countries. *Scalar invariance* requires the measurement not to have only the same metric but also the same origin across countries. Thus, item interpretations are not biased across countries. To validly compare means as well as for sophisticated analyses making use of scale

scores across countries (e.g., structural equation modeling with mean structures and multilevel analysis), scalar invariance is required.

Conventionally, multigroup confirmatory factor analysis (MGCFA) is used to check for measurement invariance. Starting with the configural model without parameter restrictions across countries, loadings (metric invariance) and intercepts (scalar invariance) are fixed to be equal across groups stepwise, while assessing change in model fit (e.g., Brown 2015). Yet, assuming identical loadings and intercepts has been criticized to be unrealistic with several countries, often leading to a poor fitting model (Muthén and Asparouhov 2014, 2018). Consequently, more flexible methods are becoming increasingly popular, constraining only a subset of parameters to be invariant (e.g., partial invariance, see Byrne et al. 1989), allowing small cross-country parameter differences (e.g., Bayesian approximate invariance testing, see B.O. Muthén and Asparouhov 2012), or favoring a model with most invariant and a minimum of non-invariant parameters (e.g., Alignment, see Asparouhov and Muthén 2014).

1.3 Empirical evidence on invariance of teaching quality measures

While the advanced methods mentioned above are highly useful from a conceptual point of view, cross-country invariance of teaching quality measures has mostly been tested by applying traditional MGCFA.

For *student support*, configural and metric but not scalar invariance have been demonstrated across many countries participating in large-scale studies (PISA 2000: Schulz 2003; TALIS 2008 2013: He and Kubacka 2015; TIMSS 2011: Nilsen and Gustafsson 2016). Likewise, *classroom management* measures satisfied configural and metric invariance but showed insufficient model fit indices for scalar invariance (PISA 2000: Schulz 2003; PISA 2012: He et al. 2017; van de Grift 2014; TALIS 2008 2013: He and Kubacka 2015; Desa 2014; TIMSS 2011: Nilsen and Gustafsson 2016). For *cognitive activation*, measurement invariance testing is scarce and has been conducted for some aspects only. Again, configural and metric but not scalar invariance were satisfied (PISA 2006 field trial data: Schulz 2005; TIMSS 2011: Nilsen and Gustafsson 2016).

The application of more flexible analysis methods is expected to fit the data more adequately and consequently yield higher levels of cross-country invariance. For instance, while not meeting scalar invariance when using MGCFA, He and Kubacka (2015) demonstrated approximate scalar invariance for classroom management measures in TALIS 2008 and 2013 using Bayesian approximate invariance testing. To our knowledge, this is the only study applying an advanced statistical method to check for invariance of teaching quality measures.

The selection of countries under investigation can also impact the degree of measurement invariance. Large-scale educational assessments aim at comparing student learning across dozens of countries. Yet, the more countries are included in a

study, the smaller the shared core of a construct becomes, making it nearly impossible to achieve scalar invariance (analysis paradox, see van de Vijver 2018b). This is supported by research in the context of teaching and learning, consistently demonstrating configural and metric but not scalar invariance across many countries (e.g., Çetin 2010; Lafontaine et al. 2018; Täht and Must 2013). However, little knowledge exists on whether testing for measurement invariance across culturally similar countries might yield higher degrees of comparability.

1.4 Reasons and empirical evidence for the impact of cultural difference on measurement invariance of teaching quality measures

Culture provides a shared understanding and meaning, and is expected to influence the interpretive and response process of survey items. Not just a common cultural knowledge, but also similar school systems, teaching practices, or construct understanding by respondents, shape how items are understood, interpreted, and answered. Thus, cultural difference can shape the meaning of teaching quality measures considerably so that they do not have the same meaning in different countries (Miller et al. 2011). Language can be seen as strong indicator for cultural closeness. Language expresses, embodies, and symbolizes cultural reality (Kramsch 1998). Words reflect a stock of knowledge about the shared world within a cultural group, such as facts, common experience, or attitudes. Moreover language identifies speakers and is a symbol of cultural identity (Kramsch 1998). Thus, linguistic similarity can be used as proxy for cultural closeness.

Research is scarce as to whether cultural differences indeed play a role for measurement invariance. A first hint can be found by comparing the study by Scherer et al. (2016) to other studies (e.g., Schulz 2003, 2005). Scherer et al. (2016) found scalar invariance of teaching quality measures for three English−speaking countries (Australia, Canada, and the USA) while scalar invariance could not be confirmed in other studies assessing invariance across vastly different countries (see Section 1.3). Yet, the question remains if the result is indeed due to linguistic similarity as this has not been tested explicitly in any study. As we additionally know that particularly countries from East Asia and Latin America showed considerable different metrics and country-specific structures of educational constructs in TALIS and PISA (He and Kubacka 2015; Schulz 2003), the question arises whether measurement invariance can be achieved within those cultural clusters from East Asia or Latin America.

1.5 The present study

As described above, there is first evidence that teaching quality measures differ across countries. However, except for Scherer et al. (2016), no study has investigated measurement invariance for the three dimensions simultaneously and findings are based on often criticized traditional analysis methods. Additionally, cultural closeness assessed via linguistic similarity seems to play a crucial role for measurement invariance and therefore needs to be included in a systematic way.

Thus, we first aim to assess the degree of cross-country invariance of items

measuring student support, classroom management, and cognitive activation using a more sophisticated method, namely the alignment optimization (Asparouhov and Muthén 2014). We hypothesize to find approximate scalar measurement invariance using that method (Hypothesis 1).

Secondly, we aim at comparing the degree of invariance of teaching quality measures across linguistically diverse countries versus linguistically similar countries. We assume to find a larger degree of measurement invariance for linguistically similar countries compared to a set of linguistically diverse countries (Hypothesis 2).

2 Method

2.1 Database and sample

The Program for International Student Assessment (PISA) 2012 survey provides data on individual students' perceptions of the three teaching quality dimensions in mathematics across 65 countries (OECD 2014).

To answer the research question whether linguistic similarity enhances measurement invariance, we included five linguistic clusters in the study. We selected the countries for each cluster based on the following criteria: (1) Each cluster consisted of countries with similar or identical testing language; (2) In addition to language similarity, we chose countries based on regional and cultural closeness; (3) To eliminate the effect of different sample sizes on the invariance results for within-cluster comparisons, each cluster was limited to three countries as only three German-speaking countries with sufficient sample size participated in PISA 2012. Fifteen educational systems/countries grouped into five linguistic clusters met the criteria and were included in the study: (Chinese-speaking) Macao, Shanghai, Taipei (=Chinese-speaking group); (English-speaking) Ireland, England (England and Wales), Scotland (=English-speaking group); (French-speaking) Belgium, France, (French-speaking), Switzerland (=French-speaking group); Austria, Germany, (German-speaking) Switzerland (=German-speaking group); Chile, Colombia, and Mexico (=Spanish-speaking group).[1] In the following, we treat all educational systems as countries for simplicity.[2]

Students with missing data on all items measuring the three dimensions of teaching quality were excluded from analysis. To avoid different model contributions due to varying sample sizes, a subsample of 1000 students per country was drawn according to final student weights (W_FSTUWT), resulting in 3000 students per linguistic cluster and a total of 15,000 students.

[1] Spain was not chosen since there are five different language versions for the autonomous Spanish communities (OECD 2014).

[2] In PISA 2012, China was represented through separate educational systems. Hong Kong was not chosen for our study, since the language of instruction is English for a major part of the student population (OECD 2014). Since Shanghai, Macao, and Taipei were treated as separate educational systems in PISA 2012, we treat them as "countries" in our study for simplicity, even though they should be referred to as cities/educational systems.

2.2 Measures

Student support is a 5-item measure, values of Cronbach's alpha range from .80 (German-speaking Switzerland) to .88 (Scotland), indicating good scale reliability across all countries. *Classroom management* is likewise measured by five items with Cronbach's alpha ranging from .81 (Colombia) to .92 (Taipei). Both scales are answered by a 4-point Likert scale ranging from 1 (every lesson) to 4 (never or hardly ever). *Cognitive activation* is a 9-item measure having a 4-point Likert scale that ranges from 1 (always or almost always) to 4 (never or rarely). Again, scale reliability is good across all countries (range Cronbach's alpha: .78 for (French-speaking) Switzerland to .87 in Scotland). A three-factor multi-group confirmatory factor analyses across all 15 countries supported metric invariance, indicating a universal factor structure across countries, with one factor for student support, one for classroom management, and one for cognitive activation (N = 15,000; CFI = 0.92; RMSEA = 0.06, change CFI and RMSEA from configural to metric model below .02 and .03 respectively, see Rutkowski and Svetina 2014). Since classroom management reflects how often there is, for instance, noise and disorder, high scores indicate high levels of classroom management, but low levels of support and cognitive activation (see Table 1).

2.3 Data analyses

To answer the research questions, we applied the alignment optimization by Asparouhov and Muthén (2014). Alignment identifies the optimal measurement invariance pattern while factor means are estimated without requiring full measurement invariance. First, the configural model is estimated with factor means fixed to zero and variances to one in all groups. Since loadings and intercepts are estimated freely, this is the best fitting model. In a second step, cross-country parameter restrictions are replaced by a procedure similar to rotation in an exploratory factor analysis, without compromising the fit of the configural model. In an iterative process, factor variance and mean values are estimated freely in order to minimize the total amount of non-invariance by applying the loss/simplicity function F. The difference of loadings and intercepts between every pair of groups is accumulated and scaled by the total loss function. Thus, F will be minimized with a *few large non-invariant* parameters combined with *many invariant* parameters. Upon minimizing F, factor means and variances are estimated. For every parameter, the largest invariant set of groups is identified. For each group not included in that set, the same parameter is considered to be non-invariant. To set the factor metric, the variance is fixed to one in group one. If *fixed* alignment is used, the factor mean is set to zero in the reference group, whereas *free* alignment estimates it as an additional parameter (see also Byrne and van de Vijver 2017; Davidov et al. 2014; Lomazzi 2018; Munck et al. 2017; Muthén and Asparouhov 2014, 2018).

Table 1 Items measuring the three basic dimensions of teaching quality in PISA 2012

Dimension	Item wording	Response scale
Student support	The teacher shows an interest in every student's learning. The teacher gives extra help when students need it. The teacher helps students with their learning. The teacher continues teaching until the students understand. The teacher gives students an opportunity to express opinions.	1 = Every lesson 2 = Most lessons 3 = Some lessons 4= Never or hardly ever
Classroom management	Students do not listen to what the teacher says. There is noise and disorder. The teacher has to wait a long time for students to <quiet down>. Students cannot work well. Students do not start working for a long time after the lesson begins.	
Cognitive activation	The teacher asks questions that make us reflect on the problem. The teacher gives problems that require us to think for an extended time. The teacher asks us to decide on our own procedures for solving complex problems. The teacher presents problems for which there is no immediately obvious method of solution. The teacher presents problems in different contexts so that students know whether they have understood the concepts. The teacher helps us to learn from mistakes we have made. The teacher asks us to explain how we have solved a problem. The teacher presents problems that require students to apply what they have learned to new contexts. The teacher gives problems that can be solved in several different ways.	1 = Always or almost always 2= Often 3 = Sometimes 4 = Never or rarely

Analyses were conducted using Mplus Version 7.4 (Muthén and Muthén 1998–2012). We applied the MLR estimator for parameter estimates that are robust to non-normality and non-independence of observations. TYPE = COMPLEX was used to account for the hierarchical data structure and TYPE = MIXTURE to specify groups (i.e., countries). The school ID was used as cluster-variable[3] to correct the standard errors based on the clustering effect. For the final measurement invariance models, we did not apply any weights for the following reasons: (1) Based on the random sample, senate weights are not needed. (2) Since contributions from each of the countries in the analysis are desired to be equal, using student weights would be contradictory. Since standard errors indicated a poor model fit using *free* alignment, the *fixed* estimation method was applied. Based on simulation studies, Asparouhov and Muthén (2014) recommend an upper limit of 25% non-invariance as a rule of thumb for trustworthy alignment results. Since teaching quality measures satisfy configural and metric but not scalar invariance in general (see Section 1.3), we focus on the possibility of valid cross-country mean comparisons and thus on the amount of non-invariant item intercepts. Latent means can be compared meaningfully, if less than 29% item intercepts of a scale

[3] Given the two-stage random sampling of students (stage 1) and schools (stage 2), PISA data does not provide information on the classroom level (Scherer et al. 2016).

are non-invariant (as suggested by Flake and McCoach 2018, based on simulation studies). We carried out two steps of analysis:

1) Checking for measurement invariance across *all* countries (not controlling for linguistic similarity) separately for student support, classroom management, and cognitive activation (three models).
2) Checking for measurement invariance *within* each language group for every dimension (15 models, resulting in a total of 18 models).

3 Results

We first describe evidence of non-invariance pertinent for factor loadings, followed by a more detailed description of item intercept non-invariance, which determines if means can be compared across countries validly. We compare measurement invariance across all countries (Hypothesis 1) versus within each linguistic cluster (Hypothesis 2). Besides testing these hypotheses, alignment identifies items with a high contribution to non-invariance, which will additionally be flagged.

3.1 Factor loading (non-)invariance

Table 2 shows factor loading non-invariance for the three teaching dimensions across all countries (Column 2) and for each linguistic group separately (Columns 3 to 7). Country codes shown in italics within parenthesis have a significantly non-invariant loading for the respective item. The percentage of non-invariant loadings with respect to the total number of loadings of each scale is shown in the row "Non-invariance."

For *student support*, the percentage of non-invariant factor loadings was the highest in the model for all countries (8% non-invariant factor loadings), followed by the model for the English-speaking countries (7% non-invariance). For the other linguistic groups, all factor loadings were invariant. For *classroom management*, the percentage of non-invariant factor loadings was again rather low, with non-invariant factor loadings only for the three Chinese−speaking countries (7% non-invariance) and across all countries (3% non-invariance). For *cognitive activation*, the same pattern emerged with non-invariant factor loadings only across all countries (1% non-invariance) and the Chinese-speaking countries (4% non-invariant factor loadings).

In total, we found factor loading non-invariance to be exceedingly low (approximate metric invariance is met in all models). Thus, associations (e.g., correlations) between variables can be compared across (linguistically diverse) countries validly.

Table 2 Factor loading measurement (non-)invariance for the three teaching quality dimensions

Item	All countries	German-speaking	Spanish-speaking	Chinese-speaking	English-speaking	French-speaking
Student support						
TS01	AUT GER CHE_D BEL FRA CHE_F CHL COL MEX QCN *(TAP)* *(MAC)* IRL ENG SCO	AUT GER CH-E_D	CHL COL MEX	QCN TAP MAC	IRL ENG SCO	BEL FRA CH-E_F
TS02	AUT GER CHE_D BEL FRA CHE_F CHL COL MEX QCN TAP MAC IRL ENG SCO	AUT GER CH-E_D	CHL COL MEX	QCN TAP MAC	IRL ENG SCO	BEL FRA CH-E_F
TS04	AUT GER CHE_D BEL FRA CHE_F CHL *(COL)* MEX QCN TAP MAC IRL *(ENG)* *(SCO)*	AUT GER CH-E_D	CHL COL MEX	QCN TAP MAC	*(IRL)* ENG SCO	BEL FRA CH-E_F
TS05	AUT GER CHE_D BEL FRA CHE_F CHL COL MEX QCN TAP MAC IRL ENG SCO	AUT GER CH-E_D	CHL COL MEX	QCN TAP MAC	IRL ENG SCO	BEL FRA CH-E_F
TS06	AUT GER CHE_D BEL FRA CHE_F CHL COL MEX QCN TAP MAC IRL ENG *(SCO)*	AUT GER CH-E_D	CHL COL MEX	QCN TAP MAC	IRL ENG SCO	BEL FRA CH-E_F
Non-invariance	8%	0%	0%	0%	7%	0%
Classroom management						
CM01	AUT GER CHE_D BEL FRA CHE_F CHL COL MEX QCN TAP *(MAC)* IRL ENG SCO	AUT GER CH-E_D	CHL COL MEX	QCN TAP *(MA-C)*	IRL ENG SCO	BEL FRA CH-E_F
CM02	AUT GER CHE_D BEL FRA CHE_F CHL COL MEX QCN TAP MAC IRL ENG SCO	AUT GER CH-E_D	CHL COL MEX	QCN TAP MAC	IRL ENG SCO	BEL FRA CH-E_F
CM04	AUT GER CHE_D BEL FRA CHE_F CHL COL MEX QCN TAP MAC IRL ENG SCO	AUT GER CH-E_D	CHL COL MEX	QCN TAP MAC	IRL ENG SCO	BEL FRA CH-E_F
CM05	*(AUT)* GER CHE_D BEL FRA CHE_F CHL COL MEX QCN TAP MAC IRL ENG SCO	AUT GER CH-E_D	CHL COL MEX	QCN TAP MAC	IRL ENG SCO	BEL FRA CH-E_F
CM06	AUT GER CHE_D BEL FRA CHE_F CHL COL MEX QCN TAP MAC IRL ENG SCO	AUT GER CH-E_D	CHL COL MEX	QCN TAP MAC	IRL ENG SCO	BEL FRA CH-E_F
Non-invariance	3%	0%	0%	7%	0%	0%
Cognitive activation						
CA01	AUT GER CHE_D BEL FRA CHE_F CHL COL MEX QCN TAP MAC IRL ENG SCO	AUT GER CH-E_D	CHL COL MEX	QCN TAP MAC	IRL ENG SCO	BEL FRA CH-E_F

Table 2 (continued)

Item	All countries	German-speaking	Spanish-speaking	Chinese-speaking	English-speaking	French-speaking
CA04	AUT GER CHE_D BEL FRA CHE_F CHL COL MEX (*QCN*) TAP MAC IRL ENG SCO	AUT GER CH-E_D	CHL COL MEX	*(QCN)* TAP MAC	IRL ENG SCO	BEL FRA CH-E_F
CA05	AUT GER CHE_D BEL FRA CHE_F CHL COL MEX QCN TAP MAC IRL ENG SCO	AUT GER CH-E_D	CHL COL MEX	QCN TAP MAC	IRL ENG SCO	BEL FRA CH-E_F
CA06	AUT GER CHE_D BEL FRA CHE_F CHL COL MEX QCN TAP MAC IRL ENG SCO	AUT GER CH-E_D	CHL COL MEX	QCN TAP MAC	IRL ENG SCO	BEL FRA CH-E_F
CA07	AUT GER CHE_D BEL FRA CHE_F CHL COL MEX QCN TAP MAC IRL ENG SCO	AUT GER CH-E_D	CHL COL MEX	QCN TAP MAC	IRL ENG SCO	BEL FRA CH-E_F
CA08	AUT GER CHE_D BEL FRA CHE_F CHL COL MEX QCN TAP MAC IRL ENG SCO	AUT GER CH-E_D	CHL COL MEX	QCN TAP MAC	IRL ENG SCO	BEL FRA CH-E_F
CA09	AUT GER CHE_D BEL FRA CHE_F CHL COL MEX QCN TAP MAC IRL ENG SCO	AUT GER CH-E_D	CHL COL MEX	QCN TAP MAC	IRL ENG SCO	BEL FRA CH-E_F
CA10	AUT GER CHE_D BEL FRA CHE_F CHL COL MEX QCN TAP MAC IRL ENG SCO	AUT GER CH-E_D	CHL COL MEX	QCN TAP MAC	IRL ENG SCO	BEL FRA CH-E_F
CA11	AUT GER CHE_D BEL FRA CHE_F CHL COL MEX QCN TAP MAC IRL ENG SCO	AUT GER CH-E_D	CHL COL MEX	QCN TAP MAC	IRL ENG SCO	BEL FRA CH-E_F
Non-invariance	1%	0%	0%	4%	0%	0%

Parentheses indicate non-invariant loadings for that specific group. *AUT* Austria, *CHE_D* (German-speaking) Switzerland, *GER* Germany, *BEL* (French-speaking) Belgium, *FRA* France, *CHE_F* (French-speaking) Switzerland, *CHL* Chile, *COL* Colombia, *MEX* Mexico, *MAC* Macao, *QCN* Shanghai, *TAP* Taipei, *IRL* Ireland, *ENG* England and Wales, *SCO* Scotland

3.2 Item intercept (non-)invariance

Table 3 shows item intercept non-invariance for the three teaching dimensions across all countries (Column 2) and for each linguistic group separately (Columns 3 to 7).

We found many more non-invariant intercepts than non-invariant factor loadings, a pattern that is in line with previous research checking for invariance of teaching quality measures. Intercept non-invariance varied according to dimension and was, compared to

Table 3 Item intercept measurement (non-)invariance for the three teaching quality dimensions

Item	All countries	German-speaking	Spanish-speaking	Chinese-speaking	English-speaking	French-speaking
Student support						
TS01	AUT GER CHE_D BEL FRA CHE_F *(CHL) (COL) (MEX) (QCN)* TAP *(MAC) (IRL)* ENG *(SCO)*	*(AUT)* GER CH-E_D	CHL COL MEX	QCN *(TAP)* MAC	IRL ENG SCO	BEL FRA CH-E_F
TS02	AUT GER *(CHE_D) (BEL)* FRA CHE_F *(CHL) (COL)* MEX QCN TAP MAC IRL ENG SCO	AUT GER CH-E_D	CHL COL MEX	QCN TAP MAC	IRL ENG SCO	BEL FRA CH-E_F
TS04	(*AUT) (GER) (CHE_D)* BEL *(FRA) (CHE_F)* CHL COL MEX QCN TAP *(MAC)* IRL ENG *(SCO)*	AUT GER CH-E_D	*(CHL)* COL MEX	QCN TAP MAC	*(IRL)* ENG SCO	*(BEL)* FRA CH-E_F
TS05	AUT GER CHE_D BEL FRA CHE_F CHL COL MEX QCN TAP MAC IRL ENG SCO	AUT GER CH-E_D	CHL COL MEX	QCN TAP MAC	IRL ENG SCO	BEL FRA CH-E_F
TS06	AUT GER *(CHE_D)* BEL FRA CHE_F CHL COL *(MEX)* QCN *(TAP)* MAC *(IRL) (ENG) (SCO)*	AUT GER *(CH-E_D)*	CHL COL MEX	QCN TAP MAC	IRL ENG SCO	BEL FRA CH-E_F
Non-invariance	32%	13%	7%	7%	7%	7%
Classroom management						
CM01	AUT GER CHE_D BEL FRA CHE_F CHL *(COL)* MEX *(QCN) (TAP) (MAC) (IRL)* ENG *(SCO)*	AUT *(GER)* CH-E_D	CHL COL MEX	*(QCN)* TAP MAC	IRL ENG SCO	BEL FRA CH-E_F
CM02	*(AUT) (GER) (CHE_D)* BEL FRA CHE_F CHL *(COL)* MEX QCN *(TAP) (MAC) (IRL)* ENG SCO	AUT GER CH-E_D	*(CHL)* COL MEX	QCN TAP MAC	*(IRL)* ENG SCO	BEL FRA CH-E_F
CM04	*(AUT)* GER CHE_D BEL FRA CHE_F *(CHL)* COL MEX QCN TAP MAC IRL ENG SCO	AUT GER CH-E_D	*(CHL)* COL MEX	*(QCN)* TAP MAC	*(IRL)* ENG SCO	BEL FRA CH-E_F
CM05	*(AUT) (GER) (CHE_D)* BEL FRA CHE_F CHL COL *(MEX) (QCN) (TAP) (MAC)* IRL ENG SC	AUT GER CH-E_D	CHL COL MEX	QCN TAP MAC	IRL ENG SCO	BEL FRA CH-E_F
CM06	AUT (GER) CHE_D BEL FRA *(CHE_F)* CHL COL MEX QCN *(TAP)* MAC *(IRL) (ENG) (SCO)*	AUT GER CH-E_D	CHL COL MEX	QCN TAP *(MA-C)*	IRL ENG SCO	BEL FRA CH-E_F
Non-invariance	37%	7%	13%	20%	13%	0%
Cognitive activation						
CA01	AUT GER CHE_D *(BEL)* FRA CHE_F CHL COL MEX QCN TAP MAC *(IRL)* ENG SCO	AUT GER CH-E_D	CHL *(CO-L)* MEX	QCN TAP MAC	*(IRL)* ENG SCO	BEL FRA CH-E_F

Table 3 (continued)

Item	All countries	German-speaking	Spanish-speaking	Chinese-speaking	English-speaking	French-speaking
CA04	*(AUT) (GER) (CHE_D)* BEL FRA CHE_F *(CHL) (COL)* MEX *(QCN) (TAP) (MAC) (IRL) (ENG) (SCO)*	AUT GER CH-E_D	*(CHL)* COL MEX	*(QCN)* TAP MAC	IRL ENG SCO	BEL FRA CH-E_F
CA05	*(AUT) (GER) (CHE_D) (BEL) (FRA)* CHE_F CHL COL MEX QCN TAP *(MAC) (IRL)* ENG SCO	AUT GER CH-E_D	CHL COL MEX	QCN TAP *(MA-C)*	*(IRL)* ENG SCO	BEL FRA *(CH-E_F)*
CA06	AUT GER CHE_D BEL FRA CHE_F *(CHL) (COL) (MEX)* QCN *(TAP)* MAC IRL ENG SCO	AUT GER CH-E_D	CHL *(CO-L)* MEX	QCN *(TAP)* MAC	IRL ENG SCO	BEL FRA CH-E_F
CA07	AUT GER CHE_D BEL FRA CHE_F CHL COL MEX QCN TAP *(MAC)* IRL ENG SCO	AUT GER CH-E_D	CHL COL MEX	QCN TAP *(MA-C)*	IRL ENG SCO	BEL FRA CH-E_F
CA08	*(AUT)* GER CHE_D *(BEL) (FRA) (CHE_F)* CHL COL MEX QCN *(TAP) (MAC) (IRL) (ENG) (SCO)*	AUT GER CH-E_D	CHL COL MEX	QCN TAP MAC	IRL ENG SCO	BEL FRA CH-E_F
CA09	AUT *(GER)* CHE_D BEL FRA CHE_F *(CHL) (COL)* MEX *(QCN) (TAP)* MAC *(IRL)* ENG SCO	AUT *(GER)* CH-E_D	CHL COL MEX	QCN TAP MAC	IRL ENG SCO	BEL FRA CH-E_F
CA10	AUT *(GER)* CHE_D BEL FRA CHE_F *(CHL)* COL MEX QCN TAP MAC *(IRL)* ENG *(SCO)*	AUT *(GER)* CH-E_D	CHL COL MEX	*(QCN)* TAP MAC	IRL ENG SCO	BEL FRA CH-E_F
CA11	AUT GER CHE_D BEL FRA CHE_F CHL COL MEX QCN *(TAP) (MAC)* IRL ENG SCO	AUT GER CH-E_D	CHL COL MEX	*(QCN)* TAP MAC	IRL ENG SCO	BEL FRA CH-E_F
Non-invariance	33%	7%	11%	22%	7%	4%

Parentheses indicate non-invariant intercepts for that specific group. *AUT* Austria, *CHE_D* (German-speaking) Switzerland, *GER* Germany, *BEL* (French-speaking) Belgium, *FRA* France, *CHE_F* (French-speaking) Switzerland, *CHL* Chile, *COL* Colombia, *MEX* Mexico, *MAC* Macao, *QCN* Shanghai, *TAP* Taipei, *IRL* Ireland, *ENG* England and Wales, *SCO* Scotland

the other dimensions, the lowest for *student support*. For all dimensions, the number of non-invariant intercepts was considerably lower when comparing countries belonging to the same linguistic cluster (0 to 22% non-invariance) compared to testing measurement invariance across linguistically diverse countries (32 to 37% non-invariant item intercepts). Yet, the amount of non-invariant intercepts varied across linguistic clusters, and was comparably low for the French-speaking country cluster (0 to 7% non-invariance) and rather high for the Chinese-speaking countries (7 to 22%) across all dimensions. While no clear pattern was found for the other linguistic clusters,

Ireland was the country showing a non-invariant intercept for the English-speaking country cluster throughout all models.

To summarize, the amount of non-invariant intercepts was below the upper limit of 29% non-invariant intercepts set as guideline for valid cross-country mean comparisons by Flake and McCoach (2018) for all three teaching dimensions when comparing countries belonging to the same linguistic cluster, allowing valid mean comparisons for that specific set of countries (approximate scalar invariance). However, with rather high intercept non-invariance for all dimensions, latent means cannot be compared across the 15 linguistically diverse countries (no approximate scalar invariance satisfied).

3.3 Intercept (non-)invariance according to item

In the following, we describe intercept non-invariance for specific items. We focus on the model testing measurement invariance across all countries. For every dimension, we highlight the two items with the highest and the two items with the lowest number of non-invariant intercepts.

For *student support*, items focusing on the students understanding (TS05 " The teacher continues teaching until the students understand." and Item TS02 "The teacher gives extra help when students need it.") seem to be particularly comparable across countries (with no and 4 out of 15 non-invariant intercepts, respectively, see Table 3). On the contrary, items focusing on the students' learning (TS01 "The teacher shows an interest in every student's learning." and TS04 "The teacher helps students with their learning.") seem to target different concepts with differing metrics across countries (with 7 out of 15 non-invariant intercepts for both items).

For *classroom management*, Item CM04 ("The teacher has to wait a long time for students to <quiet down>.") showed the lowest amount of non-invariant item intercepts (2 out 15 non-invariant intercepts). This is the only item included in our analyses that requires national adaptations (see the <> sign), whereas the remaining items showed a rather high amount of non-invariant intercepts (between 6 and 7 non-invariant intercepts).

For *cognitive activation*, Item CA07 ("The teacher presents problems in different contexts so that students know whether they have understood the concepts.") showed the lowest amount of non-invariant intercepts (1 out of 15 intercepts), followed by CA01 ("The teacher asks us to reflect on the problem.") and CA11 ("The teacher gives problems that can be solved in several different ways."), with 2 non-invariant intercepts, respectively. In contrast, Item CA04 ("The teacher gives problems that require us to think for an extended time.") and Item CA08 ("The teacher helps us to learn from mistakes we have made.") showed a rather high amount of non-invariant intercepts (11 and 9 out of 15 non-invariant intercepts, respectively).

4 Discussion

Some studies—including the OECD report on PISA 2012 results—compare correlations or even means across countries without testing for invariance of teaching quality measures (e.g., Caro et al. 2016; OECD 2013). We aimed to test whether this is justified as well as how linguistic similarity impacts measurement invariance of individual students' perceptions of teaching quality.

4.1 Limited invariance of teaching quality measures: possible sources and implications

At least two things can be learned from this study: First, if researchers are interested in comparing associations with other variables across countries, bias in all likelihood does not challenge the validity of interpretations, since measurement non-invariance for *factor loadings* was exceedingly low (approximate metric invariance reached). Second, even though we applied a more flexible method, the amount of non-invariant *item intercepts* was relatively high, overall, indicating a country-specific structure and metric of the teaching quality dimensions. Thus, Hypothesis 1, assuming approximate scalar invariance across all countries, could therefore not be confirmed, pointing out the importance of measurement invariance testing prior to evaluating cross-country mean differences in teaching quality.

At least two sources for the limited invariance of teaching quality measures found in our study are conceivable, namely scale characteristics (see Section 4.1.1) and respondents' cultural and linguistic background (see Section 4.1.2).

4.1.1 Scale characteristics

A first possible source regarding scale characteristics is *poor translation quality* triggering off divergent item meanings across countries and consequently challenging invariance (He and Kubacka 2015; van de Vijver and Tanzer 2004). Yet, PISA implemented rigorous translation procedures (e.g., back-translation and translation guidelines) to increase translation equivalence. In addition, countries with a common PISA testing language were advised to develop as similar questionnaires as possible. Building on a common linguistic base version, national questionnaires were created (OECD 2014). Thus, we expect the impact of poor translation quality to be rather low. This is supported by the low degree of measurement non-invariance we found within all linguistic clusters, except for the Chinese-speaking group. Yet, the Chinese-speaking group also jointly developed a linguistic base version, thus the rather high amount of non-invariance within the Chinese-speaking group is not expected to be caused by divergent translations.

A second possible source with respect to scale characteristics is a culture-specific meaning of specific terms. This can be mitigated by applying *national adaptations*. National adaptations adapt specific terms to a country's national and cultural context (van de Vijver 2018a). Our study supports the assumption that national adaptations have the potential of

increasing cross-country comparability: the only teaching quality item requiring a national adaptation showed the lowest amount of non-invariant intercepts for classroom management. Yet, national adaptions have to be applied carefully, as they can change the meaning of an item across countries (van de Vijver and Tanzer 2004). Thus, we recommend assessing if national adaptations ensure comparability or on the contrary lead to different item interpretations prior to data collection.

A third possible source with respect to scale characteristics concerns *item characteristics*, such as item length or item content. More complex items have been demonstrated to show higher response distortion, whereas shorter and simpler items are assumed to enhance cross-country comparability (Condon et al. 2006). We identified no consistent differences between non-invariant and invariant items with respect to item length (i.e., the short items for classroom management showed a high amount of non-invariant intercepts); instead, item content seemed to play a more important role with regard to cross-cultural comparability. First, items focusing on the students understanding seem to be particularly comparable across countries, whereas the concept and metrics of students' learning seems to differ across countries. Second, more complex items, involving more than one concept (e.g., showing *interest* in students *learning*), showed reduced cross-country comparability. Third, even though the classroom assessment scale involves short items, students across countries seem to have a different understanding of an orderly classroom environment as nearly all items showed a rather high amount of intercept non-variance. Fourth, ambiguous item wordings (i.e., extended time or complex problems) might increase the range of culture-specific interpretations; this assumption is supported by our study. We encourage further research to systematically analyze the effect of item content on cross-cultural measurement invariance of teaching quality items by additionally considering cultural differences in instruction.

4.1.2 Respondents linguistic and cultural background

Another possible source of non-invariance is *linguistic and cultural diversity* of respondents, which is supported by our study. Unlike across vastly different countries, we found measurement non-invariance to be much lower within our five linguistic country clusters (supporting Hypothesis 2, assuming a higher degree of invariance for linguistically similar countries). Yet, measurement non-invariance was rather high for the Chinese-speaking country cluster for classroom management and student support. Thus, by considering cultural and linguistic closeness, means can be compared across a subset of countries participating in large-scale studies. However, cultural diversity can impact measurement invariance in two ways:

Respondents' cultural variety can engender differences in measures. For instance, East Asian respondents (collectivism) tend to use middle categories in a response scale (modesty bias), whereas Western (individualism) and Latin-American respondents more often chose response scale end points (He and van de Vijver 2016). Thus, scores on a latent variable might reflect different levels of agreement and consequently lead to a shift of means (He and Kubacka 2015). To reduce the impact of culture-specific response tendencies, instruments aiming at reducing response effects can be applied, such as anchoring vignettes (see, e.g., He et

al. 2017; He and van de Vijver 2016).

Second, and even more problematical, respondents' cultural variety can engender a culture-specific construct meaning (van de Vijver and Tanzer 2004). Originally, the teaching quality dimensions were developed based on aspects relevant for high-quality teaching in German classrooms (Klieme et al. 2009). This might explain the high level of invariance for the German-speaking countries. Yet, instruments based on theories and models developed in a certain context might not be suitable in other contexts. Actually, our results indicate that existing instruments are not well-suited for comparisons across diverse countries. Thus, further research should investigate the understanding of high-quality teaching in additional countries (see, e.g., Praetorius et al. 2018a for a conceptualization of high-quality teaching for countries participating in the international TALIS-Video study).

One of the reasons why non-invariance occurred specifically for the Chinese-speaking group might be their relatively heterogeneity with regard to language, differing in Chinese characters (Mandarin vs. Shanghai dialect vs. Cantonese) (OECD 2015) and cultural background (e.g., different colonial history) (Schulz 2005).

One way to increase the cross-cultural suitability of survey instruments might be assessing a construct by a common core of invariant items complemented by culture-specific items (etic and emic approach, see, e.g., Cheung et al. 2011). Yet, a certain level of comparability has to be maintained (van de Vijver and He 2016). Since approximate scalar invariance was also hard to achieve across many countries using a more flexible method, it might be worthwhile to accommodate similarities and differences in measurement models in multiple cultural contexts. De Roover et al. (2017) introduced mixture simultaneous factor analysis to identify clusters of groups with similar factor structures (via a combination of latent class analysis and exploratory factor analysis). Cultural groups with similar measurement (e.g., metric invariance) can be clustered and subsequently comparisons can be done within each cluster.

4.2 Limitations and further directions

When interpreting the results of the study, some limitations have to be considered.

Our study demonstrated that linguistic similarity enhances measurement invariance. These results are in line with findings from Scherer et al. (2016) for three English-speaking countries. Further research should investigate if the results can be generalized for other linguistic groups as well as for other kinds of clusters compromising more than three countries (e.g., West European, Latin American, and Asian clusters). Additionally, by disentangling regional and linguistic closeness, the impact of language can more closely be investigated (e.g., by comparing USA, Canada, and Australia or Spain and Latin-American countries). If measurement invariance can likewise be achieved within those clusters, the number of countries for which valid mean comparisons are possible might be increased.

We used teaching quality measures on the individual level. As PISA does not contain classroom sampling, the data of students cannot be aggregated on the class level. This is unfortunate as the interpretation of many aspects of instruction is not only located on the individual but also on the class level (Lüdtke et al. 2009). We aimed at investigating measurement invariance on the country level, so future studies should test whether the results are the same when measuring teaching quality on the classroom level. Additionally, further research should consider a two-level analysis design (schools and students) and investigate the level of measurement invariance on the school and individual level.

Alignment is a promising new method for assessing measurement invariance. By overcoming often criticized strict restrictions of classical approaches, full measurement is not required for valid cross-country mean comparisons (Byrne and van de Vijver 2017). Thus, when testing for measurement invariance across many groups, we recommend alignment. However, being a new method, additional to a few existing simulation studies, further research is needed to fully answer how much non-invariance should be allowed to enable trustworthy cross-group comparisons. Asparouhov and Muthén (2014) suggest an overall limit of maximum 25% non-invariant item loadings *and* intercepts. In the case of all or nearly all loadings being invariant, it is comparably easy to stay below an overall limit of 25% non-invariance. In contrast, Flake and McCoach (2018) suggest a rule of thumb of maximum 29% non-invariant item intercepts for meaningful mean comparisons. Since the determination of the upper non-invariance limit influences interpretations, additional research on psychometric criteria is pivotal to draw valid conclusions.

Lastly, by applying quantitative measures, we were able to check for invariance of teaching quality measures and identify items challenging invariance. In a next step, the use of more qualitative approaches would be fruitful to isolate sources of non-invariance, and to provide information on the mechanisms of cultural difference on survey responding as well as information on cross-cultural instrument suitability (e.g., think-aloud techniques, see, e.g., Willis and Miller 2011).

5 Conclusions

According to Lee (2012), it "would not be an exaggeration to state that multinational perspectives are not properly represented in cross-national survey instruments to date." Our findings support this quote for teaching quality measures, indicating cross-country measurement differences. To enhance comparability, the cultural and linguistic background of respondents has to be considered for both instrument development and analysis. Further research is needed to identify limiting factors, to provide information on how a lack of invariance impacts both observed rank orders of countries (e.g., in the extent of teaching quality) and strength of correlations with other variables (e.g., student outcomes) (see also van de Vijver and He 2016).

References

Asparouhov, T., & Muthén, B. O. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal, 21*, 495−508.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., Klusmann, U., Krauss, S., Neubrand, M., & Tsai, Y. M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47*, 133−180.

Brophy, J. E. (2000). *Teaching*. Brussels: International Academy of Education.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research (second edition)*. New York: The Guilford Press.

Byrne, B. M., & van de Vijver, F. J. R. (2017). The maximum likelihood alignment approach to testing for approximate measurement invariance: a paradigmatic cross-cultural application. *Psicothema, 29*, 539−551.

Byrne, B. M., Shavelson, R. J., & Muthén, B. O. (1989). Testing for the equivalence of factor covariance and mean structures - the issue of partial measurement invariance. *Psychological Bulletin, 105*, 456−466.

Caro, D. H., Lenkeit, J., & Kyriakides, L. (2016). Teaching strategies and differential effectiveness across learning contexts: evidence from PISA 2012. *Studies in Educational Evaluation, 49*, 30−41.

Çetin, B. (2010). Cross-cultural structural parameter invariance on PISA 2006 student questionnaires. *Eurasian Journal of Educational Research, 38*, 71−89.

Cheung, F. M., van de Vijver, F. J. R., & Leong, F. T. L. (2011). Toward a new approach to the study of personality in culture. *The American Psychologist, 66*, 593−603.

Condon, L., Ferrando, P. J., & Demestre, J. (2006). A note on some item characteristics related to acquiescent responding. *Personality & Individual Differences, 40*, 403−407.

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology, 40*, 55−75.

De Roover, K., Vermunt, J. K., Timmerman, M. E., & Ceulemans, E. (2017). Mixture simultaneous factor analysis for capturing differences in latent variables between higher level units of multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal, 24*, 506−523.

Deci, E. L., & Ryan, R. M. (1996). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum.

Desa, D. (2014). Evaluating measurement invariance of TALIS 2013 complex scales: comparison between continuous and categorical multiple-group confirmatory factor analyses. *OECD Education Working Papers: Vol. 103*. Paris: OECD Publishing.

Dietrich, J., Dicke, A.-L., Kracke, B., & Noack, P. (2015). Teacher support and its influence on students' intrinsic value and effort: dimensional comparison effects across subjects. *Learning and Instruction, 39*, 45−54.

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Buttner, G. (2014). Student ratings of teaching quality in primary school: dimensions and prediction of student outcomes. *Learning and Instruction, 29*, 1−9.

Fischer, H. E., Labudde, P., Neumann, K., & Viiri, J. (2014). *Quality of instruction in physics: comparing Finland, Germany and Switzerland*. Münster: Waxmann.

Flake, J. K., & McCoach, D. B. (2018). An investigation of the alignment method with polytomous indicators under conditions of partial measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 25*, 56−70.

Gläser-Zikuda, M., Harring, M., & Rohlfs, C. (Eds.). (2017). *Handbuch Schulpädagogik [handbook school pedagogy]*. Stuttgart: UTB; Waxmann.

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

He, J., & Kubacka, K. (2015). Data comparability in the teaching and learning international survey (TALIS) 2008 and 2013. OECD education working papers: Vol.124. Paris: OECD Publishing.

He, J., & van de Vijver, F. J. R. (2016). Correcting for scale usage differences among Latin American countries, Portugal, and Spain in PISA. *Revista Electronica de Investigacion y Evaluacion Educativa*, 22.

He, J., Buchholz, J., & Klieme, E. (2017). Effects of anchoring vignettes on comparability and predictive validity of student self-reports in 64 cultures. *Journal of Cross-Cultural Psychology, 48*, 319−334.

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

Klusmann, U., Kunter, M., Trautwein, U., Lüdtke, O., & Baumert, J. (2008). Teachers' occupational well-being and quality of instruction: the important role of self-regulatory patterns. *Journal of Educational Psychology, 100*, 702−715.

Kounin, J. S. (1970). *Discipline and group management in classrooms*. New York: Holt, Rinehart & Winston.
Kramsch, C. (1998). *Language and culture*. Oxford: University Press.
Kuger, S., Klieme, E., Lüdtke, O., Schiepe-Tiska, A., & Reiss, K. (2017). Mathematikunterricht und Schülerleistung in der Sekundarstufe: Zur Validität von Schülerbefragungen in Schulleistungsstudien. [Mathematics achievement and student outcomes in secondary education: validity of student scores in educational studies]. *Zeitschrift für Erziehungswissenschaft, 20*, 61–98.
Kunter, M., Baumert, J., & Köller, O. (2007). Effective classroom management and the development of subject-related interest. *Learning and Instruction, 17*, 494–509.
Lafontaine, D., Dupont, V., Jaegers, D., & Schillings, P. (2018). Self-concept in reading: subcomponents, cross-cultural invariance and relationships with reading achievement in an international context (PIRLS 2011). *Submitted to Studies in Educational Evaluation.*
Lee, J. (2012). Conducting cognitive interviews in cross-national settings. *Assessment*, 21.
Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., & Reusser K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean Theorem. *Learning and Instruction, 19*, 257–537.
Lomazzi, V. (2018). Using alignment optimization to test the measurement invariance of gender role attitudes in 59 countries. *Methods, Data, Analyses: A Journal for Quantitative Methods and Survey Methodology, 12*, 77–103.
Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: how to use student ratings of classroom or school characteristics in multilevel modelling. *Contemporary Educational Psychology, 34*, 120–131.
Miller, K., Mont, D., Maitland, A., Altman, B., & Madans, J. (2011). Results of a cross-national structured cognitive interviewing protocol to test measures of disability. *Quality & Quantity, 45*, 801–815.
Munck, I., Barber, C., & Torney-Purta, J. (2017). *Measurement invariance in comparing attitudes toward immigrants among youth across Europe in 1999 and 2009: the alignment method applied to IEA CIVED and ICCS*. Sociological Methods & Research.
Muthén, B. O., & Asparouhov, T. (2012). Bayesian structural equation modeling: a more flexible representation of substantive theory. *Psychological Methods, 17*, 313–335.
Muthén, B. O., & Asparouhov, T. (2014). IRT studies of many groups: the alignment method. *Frontiers in Psychology, 5*, 978.
Muthén, B. O., & Asparouhov, T. (2018). Recent methods for the study of measurement invariance with many groups: alignment and random effects. *Sociological Methods & Research, 47*, 637–664.
Muthén, L. K., & Muthén, B.O (1998–2012). *Mplus User's Guide. Seventh Edition*. Los Angeles, CA: Muthén & Muthén.
Nilsen, T., & Gustafsson, J.-E. (2016). *Teacher quality, instructional quality and student outcomes: Relationships across countries, cohorts and time. IEA research for education, a series of in-depth analyses based on data of the International Association for the Evaluation of Educational Achievement (IEA)*. Cham: Springer International Publishing.
OECD. (2013). *PISA 2012 results: ready to learn (volume III) – students' engagement, drive and self-beliefs*. Paris: OECD Publishing.
OECD. (2014). *PISA 2012 technical report*. Paris: OECD Publishing.
OECD. (2015). *Codebook for PISA 2012 Main Study Student Questionnaire*. Paris: OECD Publishing.
Pinger, P., Rakoczy, K., Besser, M., & Klieme, E. (2017). Interplay of formative assessment and instructional quality—interactive effects on students' mathematics achievement. *Learning Environments Research, 47*, 133.
Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction, 31*, 2–12.
Praetorius, A.-K, Klieme, E., Bell, C.A., Qi, Y., Witherspoon, W., & Opfer, D. (2018a). Country conceptualizations of teaching quality in TALIS Video: Identifying similarities and differences. Paper presentation at the annual meeting of the American Educational Research Association, New York, NY.
Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018b). Generic dimensions of teaching quality: the German framework of Three Basic Dimensions. *ZDM, 47*, 97.
Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74*, 31–57.
Scherer, R., Nilsen, T., & Jansen, M. (2016). Evaluating individual students' perceptions of instructional quality: an investigation of their factor structure, measurement invariance, and relations to educational outcomes. *Frontiers in Psychology, 7*, 110.
Schulz, W. (2003). Validating questionnaire constructs in international studies: two examples from PISA 2000. In *Paper presentation at the annual meeting for the*. Chicago: American Educational Research Association.

Schulz, W. (2005). *Testing parameter invariance for questionnaire indices using confirmatory factor analysis and item response theory*. Paper presentation at the annual meeting for the American Educational Research Association, San Francisco.

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: the role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*, 454–499.

Täht, K., & Must, O. (2013). Comparability of educational achievement and learning attitudes across nations. *Educational Research and Evaluation, 19*, 19–38.

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 Vijver, F. J. R. (2018a). Capturing bias in structural equation modeling. In E. Davidov, P. Schmidt, J. Billiet, & B. Meuleman (Eds.), *Cross-cultural analysis: methods and applications*. New York: Routledge.

van de Vijver, F. J. R. (2018b). *Talk at the OECD-GESIS seminar: translating and adapting instruments in large-scale assessments*. Paris.

van de Vijver, F. J. R., & He, J. (2016). Bias assessment and prevention in non-cognitive outcome measures in PISA questionnaires. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Methodology of educational measurement and assessment. Assessing contexts of learning: an international perspective*. Cham: Springer International Publishing.

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis of comparative research*. Thousand Oaks: Sage.

van de Vijver, F. J. R., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: an overview. *Revue Européenne de Psychologie Appliqué, 54*, 119–135.

Walberg, H. J., & Paik, S. J. (2000). *Effective educational practices. Educational practices series*. Brussels: IAE.

Willis, G. B., & Miller, K. (2011). Cross-cultural cognitive interviewing: seeking comparability and enhancing understanding. *Field Methods, 23*, 331–341.

Yi, H. Y., & Lee, Y. (2017). A latent profile analysis and structural equation modeling of the instructional quality of mathematics classrooms based on the PISA 2012 results of Korea and Singapore. *Asia Pacific Education Review, 18*, 23–39.