Measuring teaching cross-culturally - the issue of measurement invariance and sources of bias

Dissertation

zur Erlangung des Doktorgrades der Philosophie

vorgelegt dem Fachbereich Psychologie und Sportwissenschaften

der Goethe-Universität Frankfurt am Main

von

Jessica Fischer

geboren am 26.09.1989 in Stade

Frankfurt am Main, Mai 2021

Erstgutachter: Prof. Dr. Eckhard Klieme

Zweitgutachterin: Prof. Dr. Miriam Hansen

Datum der Disputation: 07.07.2022

Table of Contents

Zusammenfassung

Im Kontext der Globalisierung nimmt das Interesse daran, Unterricht vergleichend zwischen Bildungssystemen der ganzen Welt zu untersuchen, kontinuierlich zu (Paine et al., 2016). Unterricht ist einer der stärksten Prädiktoren für Lernergebnisse von Schülerinnen und Schülern (Hattie, 2009). Folglich bieten internationale Vergleiche die einmalige Möglichkeit von besonders erfolgreichen Bildungssystemen zu lernen und geben Auskunft über die Generalisierbarkeit beziehungsweise über die kulturellen Variationen von Unterricht und dessen Wirksamkeit. Gleichzeitig sind sie richtungsweisend für bildungspolitische Entscheidungen (Klieme, 2020). Zur Erfassung von Unterrichtsmerkmalen aus der Perspektive der beteiligten Lehrkräfte und Schülerinnen und Schüler werden häufig Fragebögen in internationalen Schulleistungsstudien eingesetzt. Erste empirische Befunde weisen jedoch daraufhin, dass die Fragebogenskalen oftmals nicht *messinvariant* sind (z.B. Desa, 2014; He & Kubacka, 2015; Nilsen & Gustafsson, 2016). Das bedeutet, dass Unterschiede in den Messwerten zwischen Bildungssystemen nicht automatisch genuine Unterschiede im gemessenen Konstrukt, wie beispielsweise Unterschiede in der Klassenführung, reflektieren. Stattdessen entstehen diese teilweise durch nicht intendierte kulturelle Variationen im Antwortprozess *(Bias)*, beispielsweise durch kulturelle Unterschiede in der Bedeutung der Items zur Messung von Klassenführung oder durch kulturspezifische Antworttendenzen (van de Vijver & Leung, 1997). Eine fehlende Messinvarianz hat folgenreiche Konsequenzen, da valide (Mittelwerts-)Vergleiche von Unterrichtsmerkmalen zwischen Bildungssystemen nicht möglich sind und somit die umfangreichen Datensätze internationaler Studien nicht ausgeschöpft werden können (Davidov et al., 2018a). Dennoch mangelt es in der international vergleichenden Bildungsforschung bisher an empirischen Studien, die mit fortgeschrittenen Analysemethoden die Messinvarianz von Unterrichtsmerkmalen prüfen, sowie an empirisch-fundierten Erkenntnissen zu den Ursachen der oftmals fehlenden Invarianz. Mit einer Kombination aus quantitativen und qualitativen Methoden widmet sich die vorliegende Dissertation in drei Beiträgen der Aufarbeitung dieser Forschungslücke. Sie konzentriert sich auf Fragebogenskalen zur Messung von zwei generischen Unterrichtsmerkmalen aus der Perspektive von Schülerinnen und Schülern, der *Unterrichtsqualität* mit den Dimensionen Klassenführung, konstruktive Unterstützung und kognitive

Aktivierung und den *Unterrichtsmethoden* mit den Dimensionen lehrerzentrierte und schülerzentrierte Methoden und Methoden des Assessments.

Beitrag I prüft die Messinvarianz von PISA Skalen zur Erfassung der drei Basisdimensionen der Unterrichtsqualität zwischen 15 Bildungssystemen. Zusätzlich wird untersucht, ob die kulturelle Ähnlichkeit (operationalisiert als ähnliche oder identische Sprache) der Bildungssysteme einen Einfluss auf das Ausmaß der Messinvarianz besitzt. Da die Modellannahmen der häufig eingesetzten konfirmatorischen Faktorenanalyse zunehmend als zu strikt für Messinvarianzprüfungen im interkulturellen Kontext kritisiert werden (Rutkowski & Svetina, 2014), wird mit Alignment (Asparouhov & Muthén, 2014) eine flexiblere und angemessenere Methode verwendet. Dennoch erreichen die drei Basisdimensionen nur metrische (identische Faktorenladungen) und nicht skalare Invarianz (identische Intercepts) zwischen den 15 Bildungssystemen. Folglich sind valide Vergleiche von Mittelwertsunterschieden in der Unterrichtsqualität zwischen den 15 Bildungssystemen nicht möglich. Innerhalb der fünf Cluster, bestehend aus jeweils drei Bildungssystemen mit ähnlicher oder identischer Sprache, wird im Gegensatz dazu skalare Invarianz bestätigt. Die Ergebnisse aus Beitrag I legen nahe, dass die untersuchten Fragebogenskalen zur Messung von Unterrichtsqualität unterschiedlich zwischen Bildungssystemen funktionieren. Eine höhere Vergleichbarkeit scheint jedoch mit einer kulturellen und sprachlichen Ähnlichkeit der Befragten einherzugehen. Wird diese Ähnlichkeit bei der Analyse berücksichtigt, sind valide Vergleiche von Mittelwertsunterschieden für eine Teilmenge an Bildungssystemen mit invarianter Messung möglich.

Beitrag II knüpft an Ergebnisse aus Beitrag I an und untersucht potenzielle Ursachen der fehlenden Invarianz. Der Fokus liegt auf kulturellen Variationen im Antwortprozess, die zu einer eingeschränkten Datenvergleichbarkeit führen können (z.B. Schwarz et al., 2010). Beitrag II konzentriert sich auf die erste und zweite Stufe des Antwortprozesses, der Item-Interpretation und der Assoziation des Item-Inhaltes mit persönlichen Erfahrungen (Tourangeau, 1984). Mit Hilfe von kognitiven Interviews wird untersucht, wie Schülerinnen und Schüler aus China (Shanghai) und Deutschland PISA Items zur Messung konstruktiver Unterstützung interpretieren und welche Unterrichtserfahrungen sie mit den

Items assoziieren. Die Ergebnisse der strukturierenden qualitativen Inhaltanalyse nach Kuckartz (2018) zeigen zwar, dass sowohl chinesische als auch deutsche Schülerinnen und Schüler die Items mehrheitlich mit Unterrichtsmethoden assoziieren, die zur Kompetenzunterstützung beitragen (beispielsweise Methoden zur Beseitigung von Verständnisproblemen). Es zeigen sich jedoch auch deutliche interpretative Variationen, sowohl für statistisch nicht messinvariante (nicht vergleichbare) Items als auch für messinvariante (vergleichbare) Items. Diese können zum einen auf Eigenschaften der Messung zurückgeführt werden. Hierzu zählt eine unterschiedliche Übersetzung des Terms *Lernen* (in Deutschland *Lernfortschritt* in China *Lernstand)*. Zudem finden sich Hinweise, dass komplexe und uneindeutige Itemformulierungen mehr Spielraum für kulturspezifische Interpretationen zulassen. Die zweite Ursache der interpretativen Variationen ist ein unterschiedliches Verständnis von konstruktiver Unterstützung, das durch kulturelle Unterschiede in der Unterrichtsgestaltung und -zielsetzung erklärt werden kann (Leung, 2001). Neben der Kompetenzunterstützung assoziieren die deutschen Schülerinnen und Schüler die Items mehrheitlich mit Methoden zur Unterstützung ihrer Autonomie und ihres sozial-emotionalen Erlebens im Unterricht, wohingegen die chinesischen Schülerinnen und Schüler die Items mehrheitlich mit Methoden zur Unterstützung ihrer akademischen Produktivität (z.B. ihrer Aufmerksamkeit) assoziieren. Die Ergebnisse aus Beitrag II legen nahe, dass die Interpretation von Fragebogenitems variieren kann, je nach dem in welchem kulturellen Kontext die Frage gestellt wird. Sie betonen zudem, dass quantitative und qualitative Methoden miteinander kombiniert werden sollten, um verlässliche Information über die interkulturelle Vergleichbarkeit von Fragebogenitems zu erhalten.

Beitrag III untersucht ebenfalls potenzielle Ursachen einer fehlenden Messinvarianz und konzentriert sich dabei auf kulturelle Variationen von Unterricht. Die Operationalisierung von Unterrichtsmerkmalen in internationalen Studien basiert häufig auf theoretischen Annahmen, die für Unterricht in einem bestimmten kulturellen Kontext entwickelt wurden (z.B. Stringfield & Slavin, 1992). Es ist jedoch fraglich, ob die der Messung zugrundeliegenden theoretischen Annahmen generalisierbar sind oder ob Unterricht zwischen Bildungssystemen variiert, was die interkulturelle Passung der

Messinstrumente einschränken kann. Folglich untersucht Beitrag III für PISA Skalen, ob die theoretische Differenzierung zwischen lehrer- und schülerzentrierten Unterrichtsmethoden in zwölf Bildungssystemen empirisch bestätigt werden kann. Zudem werden Gemeinsamkeiten und Unterschiede in der Struktur und Dynamik von Unterrichtsmethoden exploriert. Die Ergebnisse der konfirmatorischen Faktorenanalyse bestätigen metrische Invarianz für Items zur Messung von lehrer- und schülerzentrierten Unterrichtsmethoden. Die theoretische Differenzierung ist demnach zwischen den zwölf untersuchten Bildungssystemen generalisierbar. Skalare Invarianz wird nicht bestätigt, was auf kulturelle Unterschiede in der Messung hindeutet. Ein Drei-Faktoren-Modell erreicht lediglich konfigurale Invarianz (identische Faktorenanzahl und Ladungsmuster), was daraufhin deutet, dass Assessment keine dritte vergleichbare latente Dimension von Unterrichtsmethoden ist. Die Ergebnisse der Netzwerkanalysen zeigen, dass sich die Netzwerke bestehend aus schülerzentrierten, lehrerzentrierten und individuellen Assessment-Methoden, in ihrer Konnektivität und globalen Struktur signifikant zwischen den untersuchten Bildungssystemen unterscheiden. Das bedeutet, dass sich die Bildungssysteme nicht nur größtenteils darin unterscheiden *welche* Methoden im Unterricht miteinander kombiniert werden, sondern auch wie *häufig* individuelle Methoden miteinander kombiniert werden. Die Ergebnisse liefern zudem Hinweise darauf, dass die Struktur von Unterrichtsmethoden vergleichbarer ist für Bildungssysteme, die einen ähnlichen sprachlichen Hintergrund besitzen (z.B. drei Bildungssysteme mit englischer Sprache), als für solche mit unterschiedlicher Sprache (z.B. Bildungssysteme mit englischer und chinesischer Sprache). Mit Blick auf die relative Wichtigkeit einzelner Methoden innerhalb der Netzwerke zeigt sich, dass Feedback zu individuellen Stärken und Schwächen aus Sicht der Schülerinnen und Schüler das Herz von Unterricht darstellt, wohingegen die Wichtigkeit der restlichen Methoden zwischen den untersuchten Bildungssystemen variiert. Die Ergebnisse aus Beitrag III legen nahe, dass, neben einigen Gemeinsamkeiten, Unterrichtprozesse zwischen Bildungssystemen variieren können. Um die interkulturelle Passung von Fragebogenskalen zu erhöhen, sollten diese Variationen bei der Operationalisierung von Unterrichtsmerkmalen berücksichtigt werden. Die Ergebnisse betonen

zudem, dass die Grundlage internationaler Studien eine gemeinsame Konzeptualisierung sein sollte, anstelle von Theorien, die sich auf einen bestimmten kulturellen Kontext beziehen.

Die vorliegende Arbeit leistet einen wesentlichen Beitrag zu einer vergleichbaren und validen Messung von Unterrichtsmerkmalen im interkulturellen Kontext. Durch die Kombination aus quantitativen und qualitativen Methoden identifiziert sie nicht nur Bildungssysteme, Konstrukte und Fragebogenitems, für die eine interkulturelle Messung besonders problematisch ist, sondern ermittelt gleichzeitig Ursachen der fehlenden Vergleichbarkeit. Sie zeigt, dass die kulturelle Heterogenität der untersuchten Bildungssysteme die Vergleichbarkeit von Fragebogenskalen auf vielfältige Weise beeinträchtigen kann. Hierzu gehören kulturelle Variationen in der gesprochenen Sprache, der Gestaltung von Unterricht oder der kognitiven Informationsverarbeitung, die zu systematischen Unterschieden in der Fragebogenbeantwortung führen können. Die Arbeit kritisiert demnach einen "one-size-fits-all pedagogical approach" (Tabulawa, 2003, p.9) vieler interkultureller Studien als unzureichend, um der kulturellen Vielfalt der teilnehmenden Bildungssysteme gerecht zu werden. Stattdessen gibt die Arbeit auf Basis der Ergebnisse ihres Mixed-Methods-Ansatzes spezifische Empfehlungen, wie die Berücksichtigung kultureller Variationen in verschiedenen Stadien einer interkulturellen Studie zu einer validen und vergleichbaren Messung und Analyse von interkulturellen Daten zu Unterrichtsmerkmalen beitragen kann.

1 The challenge of measuring teaching cross-culturally

1.1 Conceptualizing something as complex as teaching

Around the world, teaching has a significant and unparalleled impact on cognitive (e.g., knowledge) and non-cognitive student learning outcomes (e.g., goal orientation, Hattie, 2009). Moreover, in contrast to other factors relevant for student learning (e.g., the student's socio-economic background), teaching is more malleable and thus can be subjected to targeted interventions (OECD, 2013a). Yet, teaching is highly complex and comprises a wide range of aspects that occur in varied settings (Praetorius et al., 2019). Moreover, teaching is never a linear process, instead it has a dynamic nature and many teaching practices typically occur simultaneously, making it difficult to disentangle practices and their effects (Opfer et al., 2020). Consequently, there are a wealth of ways for conceptualizing teaching. Based on the Input-Process-Product model (Purves, 1987), educational psychology often conceptualizes teaching *processes* as a complex interaction of three components which have an impact on students' learning outcomes: teaching quality, teaching practices, and teaching content (i.e., opportunity to learn) (Kuger et al., 2017). This dissertation focuses on the two generic components of teaching, i.e., teaching quality and teaching practices[1].

Teaching quality describes the quality of the interactions between the students and the teacher (i.e., the "deep structures" of teaching, Kunter & Ewald, 2016). It has been conceptualized in various frameworks (for an overview see Charalambous & Praetorius, 2018). One prominent framework is the framework of the Three Basic Dimensions, comprising the dimensions of classroom management, student support, and cognitive activation (Klieme et al., 2009). *Classroom management* refers to a range of practices and structures that help teachers to maximise learning time, such as the effective handling of disruptions, use of time, routines, and monitoring. *Student support* encompasses practices with a focus on creating classroom environments in which students can learn and develop, such as supporting social relationships, autonomy, and competence. *Cognitive activation* comprises teaching

[1] For the sake of simplicity and easier reading, in the following "teaching" refers to teaching quality and teaching practices.

aspects that promote students' higher-level thinking, such as the use of challenging tasks, exploring and activating prior knowledge, supporting metacognition, or eliciting student thinking (Praetorius et al., 2018b). The three dimensions have been positively linked to student cognitive and non-cognitive outcomes (e.g., Baumert et al., 2010; Caro et al., 2016; Hattie, 2009; Pianta et al., 2012).

Teaching practices describe actual teaching methods (i.e., the "visual structures" of teaching, Kunter & Ewald, 2016). The conceptualization of teaching practices has been inspired by two predominant philosophies of education that evolved decades ago and guide designs of instruction to date, namely instructionism (based on Rosenshine, 1976) and constructivism (based on Dewey, 1929; Piaget, 1952; Vygotsky, 1978). Instructionism advocates the use of *teacher-directed practices* with the aim to provide a well-structured and effective learning environment. The teacher is the transmitter of knowledge and controls the learning processes in the classroom (Caro et al., 2016). *Student-centred practices* based on constructivism foster students' active engagement in their learning processes and promote a self-directed construction of knowledge while the teacher supports and guides the learning processes (Tobias & Duffy, 2009). More recently, *classroom assessment* as an additional element of teaching practices has garnered much attention (OECD, 2013b). Classroom assessment practices are used to evaluate students' knowledge and progress and can either serve the purpose of summarizing achievement (Harlen & Deakin-Crick, 2002) or the formative purpose of improving teaching and learning according to an ongoing assessment basis that includes the provision of feedback (Black & Wiliam, 2009). While teacher-directed and student-centred practices represent traditional (instructionism) and modern (constructivism) approaches to instruction, it remains unclear how classroom assessment integrates into the teaching practice framework. A combination of specific practices (e.g., feedback and formative assessment, Hattie, 2009) and well-implemented teacher-directed practices have been positively associated with achievement, while the relationship between student-centred practices and achievement showed inconsistent results (Caro et al., 2016) and the mere frequency of applied practices seems to have little effect on student outcomes (Klieme et al., 2010).

To date, the relationship between teaching quality and teaching practices remains vague. Specific practices may be linked to individual dimensions of teaching quality. For instance, teacher-directed instruction may show a higher degree of structuring compared to student-centred instruction in general. Still, the mere implementation of specific practices does not automatically result in high-quality teaching (Kunter & Ewald, 2016).

In the context of globalization, the perspective on teaching has shifted from a national to an international level: "Teaching […] today means that discussions of teaching are no longer local ones, nor are they understood solely in national terms" instead "[…] teaching is considered through the lens of globalization – that is, understanding teaching in one place in relation to teaching in other places and defining the goals of teaching in part by a vision of a changing and more interconnected world" (Paine et al., 2016, p. 717). Consequently, the interest in studying teaching on an international level has steadily increased over recent decades, which is mirrored by a variety of books that have been published in the field of cross-cultural educational psychology[2]. A shared goal is to gain insights into the variation and generalizability of teaching and to provide indicators of education systems' effectiveness, equity, and efficiency, which can help to inform educational policy on a regional, national, and international level. A prime resource to address these goals are international large-scale assessments (ILSAs). Currently, the most frequently cited ILSAs include the Programme for International Student Assessment (PISA), the Trends in International Mathematics and Science Study (TIMSS), and the Teaching and Learning International Survey (TALIS). Besides providing indicators of students' achievement, ILSAs measure pivotal factors contributing to learning outcomes such as teaching quality and practices across dozens of education systems in so-called "context-

[2] For instance:
The SAGE handbook on comparative studies in education (Eds. L.E. Suter, E. Smith, & B.D. Denman, 2019);
International perspectives in educational effectiveness research (Eds. J. Hall, A. Lindorff, & P. Sammons, 2020);
Advancing cross-cultural perspectives on educational psychology. A festschrift for D.M. McInerney (Eds. G.A.D. Liem & A.B.I. Bernardo, 2013) or
Assessing contexts of learning. Methodology of educational measurement and assessment (Eds. S. Kuger, E. Klieme, N. Jude, & D. Kaplan, 2016).

questionnaires" (Klieme, 2020). Yet, as Opfer and colleagues (2020, p.6) concluded: "If measuring something as complex and context based as teaching is challenging, doing so […] at an international scale is even harder". The methodological challenge to validly measure teaching across education systems[3] via (student) questionnaires is focused by this dissertation. Particular emphasis is placed on multiple-item measures with a Likert-type response format. Yet, to meaningfully describe these challenges (Section 1.4) and derive the dissertation aims (Section 2), the concepts of Bias and Invariance are first introduced, their being fundamental to cross-cultural research.

1.2 The framework of bias and measurement invariance

When teaching is measured cross-culturally, the validity of questionnaire measures may seriously be compromised by several nuisance factors which are specifically caused by the intercultural context. In the framework of Bias and Invariance (also termed equivalence) (van de Vijver & Leung, 1997), these nuisance factors are referred to as ***Bias*****.** If a questionnaire measure is biased, scores from the assessment do not reflect genuine cross-cultural differences in the targeted construct but are rather partially caused by unintended systematic cultural variations affecting survey responses. For instance, individual differences in achievement test scores may reflect differences in intelligence in a single cultural group, whereas intergroup differences may be largely due to culture-specific differences in education and test experience (van de Vijver, 1998). Depending on the source of variations, three types of bias can be distinguished. *Construct bias* is characterized by a different meaning of the targeted construct across cultural groups; *method bias* is induced by cross-cultural variations in methodological-procedural aspects of a study, such as sampling, use of the questionnaire measures (e.g., culture-specific response tendencies), or administration modes; and *item bias* (also differential item functioning DIF) is triggered by a different meaning of items across cultural groups (van de Vijver, 1998; van de Vijver & Leung, 1997). Thus, if there is bias, there is no way of validly comparing data obtained

[3] In the following, I refer to education systems instead of countries as some countries (e.g., China) have more than one education system and differences in education systems may have a bearing on the measurement of teaching.

in different education systems. In order to avoid erroneous or ambiguous inferences based on biased data, preventive measures should be taken to minimize bias and enhance comparability at various stages of a cross-cultural study (for an overview of preventive measures see van de Vijver & He, 2016).

Even when carefully designing and administrating cross-cultural studies, bias might still occur to some extent. After data collection, checking ***measurement invariance*** is the ultimate gatekeeper to ensuring valid cross-cultural comparisons. Measurement invariance refers to whether and at which level scores obtained from different education systems can be validly compared. Traditionally, at least three hierarchically linked levels can be distinguished, and implications are attached to each level of invariance reached (van de Vijver & Leung, 1997). *Configural invariance* indicates that items measuring a construct cover facets of that construct adequately in all education systems studied. Hence, the basic structure of a construct can be studied comparatively (Vandenberg & Lance, 2000). *Metric invariance* implies that the construct is not only measured with the same items, but also with the same metric. This means that across education systems, each item of a scale is equally related to the targeted construct. Thus, an increase in a survey response to an item is associated with the same increase in that construct (Fischer & Karl, 2019). If metric invariance is reached, correlates of the measured construct (covariances and unstandardized regression coefficients) can be compared (Steenkamp & Baumgartner, 1998). *Scalar invariance* requires the measurement to have both the same metric and the same origin across education systems. Only with scalar invariance are scale scores not biased and items are understood and answered in the same way regardless of group membership. Consequently, latent mean scores can be compared validly (Marsh et al., 2009) and sophisticated analyses are possible that make use of scale scores (e.g., multivariate analysis or structural equation modeling with mean structures, Davidov et al., 2014).

If measurement invariance is not tested, comparative research on teaching runs the risk of drawing erroneous inferences. Structural relations between constructs (e.g., student support and achievement) may be masked, exaggerated, or mainly due to measurement artefacts (Davidov et al., 2018a). Moreover, the computation of incorrect group means may result in equivocal rankings of education

systems (Little, 2013) and cross-cultural educational research may misinform and mislead educational policy (Klieme, 2020). Thus, to avoid erroneous conclusions, it is vital to statistically check the invariance of teaching measures before drawing any cross-cultural comparisons of similarities and differences in teaching.

1.3 Conventional method to check measurement invariance: multigroup confirmatory factor analysis

Various psychometric methods are available to check measurement invariance, ***Multigroup confirmatory factor analysis (MGCFA,*** Jöreskog, 1971) being the most rigorous and widely used. The underlying assumption is that items are indicators of latent factors and responses to these items are "caused" by the latent factors. Based on the covariance matrix information, a series of hierarchical models is tested which are linked to the three measurement invariance levels. First, the *configural model* without cross-group parameter constraints is estimated. It requires that the mere factor structure is equal across groups, meaning the same number of factors and items that load on each factor is found (Davidov et al., 2014). Afterwards, factor loadings *(metric model)* and subsequently item intercepts *(scalar model)* are fixed to be equal across groups. The level of invariance is inferred from the fit indices in each model and by comparing model fit measures between the more and less constrained models (e.g., metric versus configural model). Commonly used fit indices include the Comparative Fit Index (CFI), the Root Mean Square Error of Approximation (RMSEA), or the Standardized Root Mean Residual (SRMR) along with Chi-square statistics (Cheung & Rensvold, 2002; Hu & Bentler, 1999; Rutkowski & Svetina, 2017; for comparisons of many groups see Rutkowski & Svetina, 2014).

1.4 The unachievable ideal: scalar non-invariance of teaching measures

Despite its critical relevance, measurement invariance testing is only just evolving in the field of cross-cultural educational psychology with regard to non-cognitive constructs (Klieme, 2020). Initial studies have applied MGCFA to check measurement invariance of teaching quality and practice measures, particularly for education systems that participated in ILSAs.

For measures of c*lassroom management, student support,* and *cognitive activation,* configural and metric but not scalar invariance have been demonstrated across multiple education systems (TALIS 2008, 2013: Desa, 2014; He & Kubacka, 2015; Global Teaching InSights: a video study of teaching (GTI): Mihaly et al., 2021; TIMSS 2011: Nilsen & Gustafsson, 2016). In some studies, measures of classroom management reached configural invariance only (TALIS 2018: OECD, 2019; GTI: Mihaly et al., 2021). Likewise, across multiple education systems, measures of *teacher-directed* and *student-centred practices* satisfied configural and metric but not scalar invariance (TALIS 2008, 2013: Desa, 2014; He & Kubacka, 2015). For *classroom assessment* again configural and metric invariance have been established, unlike scalar invariance (PISA 2015: Klieme, 2020).

Thus, *if* measurement invariance has been tested, findings consistently point towards a lack of scalar invariance of teaching measures. Hence, the questionnaire measures under investigation seem to work differently across education systems: while the items seem to represent the same unidimensional latent construct (configural invariance) and have a similar discriminant power (metric invariance), the item difficulties seem to vary across the compared education systems (no scalar invariance). These findings of scalar non-invariance are consistent with measurement invariance investigations in the field of cross-cultural educational psychology for additional constructs (e.g., Çetin, 2010; He et al., 2019; Lafontaine et al., 2019; Täht & Must, 2013; van de Grift, 2014).

Returning to the statement by Opfer and colleagues (2020, p.6): "If measuring something as complex and context-based as teaching is challenging, doing so […] at an international scale is even harder" - the lack of (scalar) invariance is probably one of the biggest challenges in cross-cultural educational psychology. Scalar non-invariance is particularly troublesome as it prevents researchers from conducting meaningful cross-cultural comparisons of mean differences in teaching. On the other hand, if researchers refrain from any comparisons due to (scalar) non-invariance, this might mean that the wealth of data from cross-cultural studies are not exhausted and their potential is not realized (Davidov et al., 2018a).

2 Dissertation aims

Given the methodological challenge faced by cross-cultural educational psychology, the overarching goal of this cumulative dissertation is to identify strategies for dealing with the limited (scalar) invariance of Likert-type questionnaire measures of teaching quality and practices, and to contribute to a more comparable cross-cultural measurement of teaching. A focus is placed on the advancement of methods and strategies of measurement invariance testing in the field of cross-cultural educational psychology *(Dissertation aims 1 and 2)*, with an additional aim to understand why invariance of teaching measures is often not given (*Dissertation aims 3 and 4).* To address the research aims empirically, this dissertation draws on various quantitative and qualitative methods with complementary strengths. In the following, the dissertation aims are derived from relevant theoretical considerations and empirical research, and the methods to address these research aims are introduced. Afterwards, a summary of the three manuscripts that address the dissertation aims empirically is provided (Section 3).

2.1 Advancement of methods and strategies of measurement invariance testing

2.1.1 A sophisticated method to check measurement invariance: the alignment method

Despite its prevalence in the field of cross-cultural psychology, MGCFA has been called unsuited for investigating measurement invariance across multiple education systems for several reasons: First, MGCFA was initially developed for comparisons of two groups, thus the model fit criteria may not apply to comparisons involving many groups (Rutkowski & Svetina, 2014). Further, it has been argued that the assumption of identical factor loadings and item intercepts is overly strict (Muthén & Asparouhov, 2012). Consequently, scalar invariance rarely fits the data well and can be described as an unachievable ideal in a cross-cultural context (Marsh et al., 2018). Finally, it is nearly impossible to determine whether non-invariance is caused by model misspecifications with severe consequences for comparability or from minor misspecifications that still allow meaningful comparisons (Oberski, 2014). To overcome these limitations, more advanced methods have been proposed (see Table 1). The common assumption is that small violations of invariance are not consequential for meaningful cross-cultural comparisons (Davidov et al., 2014).

Table 1 Advanced methods for measurement invariance testing (most widely used Structural Equation Modeling (SEM)-related approaches)

Approach	Advantages (compared to MGCFA) and limitations	Reference
Partial invariance (in MGCFA)	+ Allows a subset of factor loadings and item intercepts to vary if large deviations of item functions are detected while keeping others invariant - No clear guideline on the amount of parameters to free	Byrne et al. (1989)
Exploratory Structural Equation Modeling (ESEM)	+ Factor loadings are freely estimated, relaxes all zero-cross-loadings to find the model that fits the data best + Performs well with many cross-loadings and may better represent substantive theories - Too many cross-loadings and multiple corrections can invalidate the measurement model	Asparouhov & Muthén (2009)
Bayesian structural equation modeling (BSEM)	+ Permits “wiggle room”, i.e., small differences across groups in factor loadings and intercepts (approximate invariance) + Researchers may incorporate their prior knowledge on parameters + May better represent substantive theories - Several models with different prior variances are needed to identify best model - No clear model evaluation criteria - With increasing prior variance and sample size, the execution time increases	Muthén & Asparouhov (2012)
Alignment	+ Searches for an optimal solution which entails only few parameters with large differences across groups + Relaxes parameters that are very different across groups while keeping others invariant + Easy to use and performs well for many groups + Provides detailed information on non-invariance and factor mean differences + May better represent substantive theories - Conditional on configural invariance - Difficult to ensure the simplest model is the “correct” model - No clear guideline about how much non-invariance is permissible for valid comparisons	Asparouhov & Muthén (2014)

Particularly the ***alignment method*** by Asparouhov & Muthén (2014) is a promising alternative to MGCFA. Alignment only needs two steps to test measurement invariance. First, the *configural model* is estimated with free factor loadings and intercepts across groups and fixed factor means and variances. As this model imposes no cross-group parameter constraints, it is the best fitting model. Afterwards, factor means and variances are estimated freely without compromising the fit of the configural model. Their values are chosen by using a simplicity function which minimizes parameter differences across groups (instead of constraining them to be equal) and is similar to rotation in exploratory factor analysis. The simplicity function is optimized at a few large non-invariant parameters and many approximately invariant parameters. When the total amount of non-invariant parameters is minimized, measurement parameters and factor means are estimated for each group (Asparouhov & Muthén, 2014). As these estimated means take detected differences between parameters into account, they provide the best possible comparability that can be achieved with the given data (van de Vijver et al., 2019). Two further technical details are worth noting. Alignment can be based on maximum likelihood or Bayes estimation and allows two estimation options: With *free* alignment latent means are estimated freely, and with *fixed* alignment the latent mean is fixed to zero for the reference group (Asparouhov & Muthén, 2014; Davidov et al., 2018b; Muthén & Asparouhov, 2014). Asparouhov and Muthén (2014) recommend an upper limit of 25% non-invariant factor loadings and item intercepts to compare means validly (based on simulation studies). Moreover, alignment identifies non-invariant parameters for each group, provides the factor mean ordering among groups and significant differences, and information on each items' intercept and loading contribution to the optimized simplicity function (Muthén & Asparouhov, 2018).

In a cross-cultural context, alignment has several advantages over MGCFA: First, alignment estimates group-specific factor means and variances without requiring full scalar invariance. Moreover, simulation studies have demonstrated that it performs well for comparisons of two as well as across many groups and that only in extreme circumstances with a large amount of non-invariant parameters, the model estimates may be biased (e.g., Asparouhov & Muthén, 2014; Flake & McCoach,

2018; Muthén & Asparouhov, 2018; Robitzsch, 2020). Finally, alignment is expected to fit cross-cultural data more adequately and to yield higher levels of invariance, and therefore it may allow valid mean comparisons even if full measurement invariance is not given (Byrne & van de Vijver, 2017). A first hint supporting these assumptions is provided by He and Kubacka (2015). While scalar invariance was not established across all education systems that participated in TALIS 2008 with MGCFA, the authors demonstrated approximate scalar invariance with Bayesian invariance testing for classroom management. However, as elaborated in Section 1.4, the few studies that have checked measurement invariance of teaching measures so far have mostly applied MGCFA and the lack of scalar invariance may be a result of unrealistically strict assumptions (identical factor loadings and intercepts). Thus, an advancement of psychometric methods of measurement invariance testing in the field of cross-cultural educational psychology is attempted here.

Dissertation aim 1*: The first aim is to check the degree of cross-cultural invariance of teaching measures using a more sophisticated method, namely alignment (Manuscript 1)*

2.1.2 Checking measurement invariance across cultural clusters

In the field of cross-cultural educational psychology, it has been argued that the lack of scalar invariance may also be linked to the cultural heterogeneity of the compared education systems (Rutkowski & Rutkowski, 2019). According to Hwang and Matsumoto (2013, p.21): "Culture is one of the biggest challenges we face as educators and researchers because culture is an entity that involves complex social structures, elements and their meanings". Therefore, *culture* describes something that is *shared* between members of the same cultural group, such as geography, heritage and traditions, or social institutions. This shared culture is rooted in cognitive abilities and provides a meaning and an information system to its members, which involves guidelines for how to behave, think, and feel in a social context (Hwang & Matsumoto, 2013). While culture unites members of the same cultural group, it equally differentiates them from members of other cultural groups, with their own culture and meaning and information system. Consequently, depending on the cultural context, members of

different education systems[4] may differ in their ways of thinking and making sense of oneself, others, and the world, cognitive procedures, and goals (Schwarz, 2007). These cultural variations may lead to differences in survey responses across cultural groups and therefore they may compromise the invariance of survey measures. For instance, students from different cultures have been shown to differ in their understanding of items referring to their parents' wealth and possessions. While having a car may not indicate the parents' socio-economic status in the United States - as almost everyone has a car regardless of the level of income (due to large distances between locations), it may be perceived as a symbol of wealth in Japan, where having a car is less common (due to shorter distances between locations and good public transportation, Khorramdel et al., 2020). Besides cultural variations in social and geographical structures, cultural factors not relevant to the target construct may bias the cross-cultural assessment of educational constructs in many ways (see Section 2.2). Moreover, the extent to which culture may compromise the invariance of survey measures may also increase in relation to cultural distance between the compared education systems. As stated by van de Vijver and Matsumoto (2011, p.3): "The larger the cross-cultural distance between groups, the more likely cross-cultural differences will be observed, but the more likely these differences may be influenced by uncontrolled variables". Or in other words, the more culturally distant the education systems under comparison are, the more likely it is for the assessment to be biased by unintended cultural variations, and the less likely it is to achieve scalar invariance.

While some aspects are culture-specific, there are also aspects that are shared between a subset of cultures or that can even be seen as universal. For instance, Hofstede (2001) proposes that cultures can be grouped into clusters according to cultural dimensions, such as their affluence level, cultural values of individualism-collectivism, or power distance (i.e., the degree that less powerful members of institutions accept the unequal power distribution). Thus, according to their values on these

[4] In my dissertation, I target "mainstream" students and do not differentiate between members of different cultural groups within education systems.

dimensions, some cultures are more alike than others. Besides Hofstede's dimensions, another indicator of cultural closeness is a similar or identical language. Words reflect a stock of knowledge about the shared world within a cultural group, such as facts, common experience, or attitudes. Consequently, language identifies speakers and symbolizes a shared cultural identity (Kramsch, 1998). In the cross-cultural assessment of teaching, the respondents' language is often more easily accessible compared to Hofstede's dimensions. Therefore, I operationalise linguistic similarity as an indicator of cultural closeness. Returning to the assumption that cultural heterogeneity may compromise the invariance of survey measures, restricting comparisons to education systems with similar cultural background may yield higher levels of invariance. Initial empirical evidence supports this assumption. Contrary to across multiple education systems, scalar invariance was established across a subset of education systems that participated in cross-cultural studies (e.g., Davidov et al., 2008; He & Kubacka, 2015; Welkenhuysen-Gybels et al., 2007), and the lack of scalar invariance could be attributed to a few education systems with pronounced cultural differences that contrast the remaining participating education systems (He & Kubacka, 2015; Maulana et al., 2021).

However, research is still scarce as to whether cultural heterogeneity has an impact on the degree of invariance with respect to teaching measures. An initial study by Scherer et al. (2016) found scalar invariance of teaching quality measures for three English–speaking education systems, while scalar invariance could not be confirmed in other studies assessing invariance of teaching quality measures across multiple education systems (see Section 1.4). Likewise, Klieme (2020) demonstrated scalar invariance across four English-speaking education systems but not across multiple education systems with MGCFA for classroom assessment. Yet, Scherer and colleagues (2016) only tested measurement invariance across three education systems with a similar language, thus the question remains if the result is indeed due to linguistic similarity as this was not tested explicitly. Moreover, so far scalar invariance has only been achieved for English-speaking education systems (Klieme, 2020; Scherer et al., 2016). Thus, to draw substantial conclusions regarding the impact of linguistic similarity on the invariance of teaching measures, this research needs to be extended to additional clusters of education

systems, see the second aim of this thesis. This is particularly vital as considering linguistic similarity may enable valid mean comparisons of differences in teaching after all, at least for a subset of education systems with invariant measurement.

Dissertation aim 2*: The second aim is to investigate the impact of linguistic similarity on the degree of cross-cultural invariance of teaching measures (Manuscript 1)*

2.2 Understanding limited cross-cultural invariance: identification of sources of bias

The first two dissertation aims address strategies that may help overcome the often demonstrated scalar non-invariance of teaching measures, and that may enable valid mean comparisons even if full invariance is not established across multiple education systems. However, measurement non-invariance may still persist and only by understanding the root-causes of non-invariance, the invariance of teaching measures can be enhanced. As highlighted in the previous section, culture is expected to have an impact on the degree of measurement invariance. Thus, to enhance invariance, it is vital to investigate cultural variations between education systems more closely. Cultural variations that cause bias in the design stage of a study respectively cultural variations in methodological-procedural aspects of a study have been researched comparatively well[5]. Hence, this dissertation investigates cultural variations that may produce bias on the construct and item level, namely cultural variations in the structures of teaching (Section 2.2.1) and the cognitive processing of survey items (Section 2.2.2). As measurement invariance testing methods are limited in their ability to identify sources of non-invariance (Meitinger, 2017), alternative methods are introduced.

2.2.1 Cultural variations in the structures of teaching

As highlighted in the introduction, teaching has a complex and content-based nature. This intrinsic complexity is further compounded by the cultural context in which it takes place (Opfer et al., 2020): "Despite similarities in the nature of teaching as an enterprise directed toward instructing large groups

[5] For bias based on the administration mode see OECD, 2014; for sampling bias see Boehnke et al., 2010; for response tendencies see Kyllonen & Bertling, 2014 or van de Gaer et al., 2012.

of students […] there is much variation in pedagogical approaches within and particularly across countries" (Paine et al., 2016, p. 732). The characterization of teaching as *a national activity* was particularly shaped by the TIMSS Video studies, which provided initial evidence that teaching is much the same within education systems, yet much is different across education systems (Givvin et al., 2005; Stigler & Hiebert, 1999). These findings of difference have been echoed in subsequent research (e.g., Clarke & Xu, 2008; Leung, 2005; Osborn et al., 2003; Santagata, 2005; Tobin et al., 2009). Classroom space, use of time, discourse, classroom activities, and teachers' decisions on what and how to teach have been shown to vary across education systems (Alexander, 2000).

These differences in teaching may be explained by the fact that "[…] teaching is powerfully shaped by contextual factors including material conditions, institutional norms, and cultural practices and beliefs" (Paine et al., 2016, p. 732). For instance, education systems may have aligned their teaching to culturally shaped beliefs of good practices (Givvin et al., 2005). Praetorius and colleagues (2018a) surveyed educational researchers regarding what constitutes good practices in their respective education systems, and they found substantial differences with respect to the categorization of good practices depending on pedagogical traditions and national cultures. On the other hand, the effectiveness of teaching may be moderated by differences in education systems or economic and cultural factors (Kyriakides et al., 2020). Fuller and Clarke (1994) argue that student-centred practices promoting an active engagement of students during instruction are incompatible with strong hierarchical structures in cultures valuing power distance. Likewise, Alavi and McCormick (2004) postulate that practices promoting teachers' critical reflection and inquiry might be less effective in collectivist cultures, where criticism is communicated more indirectly than in Western cultures. Consequently, education systems may differ with respect to their preferred approaches to teaching and the combination of individual practices (Echazarra et al., 2016). Hence, it is vital to consider the context specificity of teaching when measuring teaching cross-culturally.

Nevertheless, teaching is often conceptualized and operationalized based on a "one-size-fits-all pedagogical approach […] that works with equal effectiveness irrespective of the context" (Tabulawa,

2003, p.9). For instance, the conceptualizations of teaching quality and teaching practices are based on well-founded theoretical considerations matched to teaching in Western education systems (see Section 1.1). Meanwhile, both constructs have been measured across multiple education systems (e.g., PISA: OECD, 2014; TALIS: Desa, 2014), often with measures that have been inspired by these Western-based frameworks and empirical evidence based on teaching in Western education systems[6]. However, given the assumption that culture and pedagogical traditions may considerably shape the designs of teaching, instruments based on theories and models developed in a certain context may not be transferable to other (non-Western) contexts (van de Vijver, 1998). Consequently, survey instruments may work differently across education systems, which can compromise the invariance of teaching measures. Yet, understanding and considering cultural differences in teaching can aid a valid measurement and inform about the cross-cultural suitability of survey instruments. Unfortunately, how classroom teaching around the world varies remains largely unexplored for teaching practices and quality (Opfer et al., 2020), and comparisons are often limited to specific education systems (Paine et al., 2016). Thus, this dissertations' third contribution is to provide explorative insights into the structure of teaching across education systems as a potential source of non-invariance of teaching measures.

Dissertation aim 3*: The third aim is to investigate sources of the limited invariance of teaching measures with regard to cultural variations in the structure of teaching (Manuscript 3)*

To gain insight into structures of teaching across education systems, it is important to consider the dynamic nature of teaching and the interrelation between individual practices, which may differ across education systems. Psychometric ***network analysis***[7] (Epskamp et al., 2018) aligns well with the theoretical assumptions of a dynamic nature of teaching and has been employed in a broad range of

[6] For instance, the PISA student support scale was developed based on work by Stringfield & Slavin (1992) based on teaching in Western education systems.

[7] Psychometric network analysis is not the same as *social* network analysis that is used to analyse the structure of social relations such as kinship structure or social mobility (Scott, 1988).

studies, particularly within the field of clinical psychology (e.g., Borsboom & Cramer, 2013; van Borkulo et al., 2015), educational (e.g., Abacioglu et al., 2019; Sachisthal et al., 2019) and personality research (e.g., Costantini et al., 2015), and research on political attitudes (e.g., Dalege et al., 2018).

While factor analytic models assume that indicators (individual items) are causally dependent on a shared latent factor, network analysis conceptualizes a construct (e.g., teacher-directed practices) as a dynamic network of indicators, which mutually reinforce one another - when one indicator changes (e.g., "the teacher sets learning goals"), so does the other, connected indicator (e.g., "the teacher tells students what they have to learn"). Thus, indicators are part of the construct instead of being measures of it (Sachisthal et al., 2019). In a network, indicators are represented by *nodes* and the unknown statistical relation between two nodes is represented by *edges*, and the magnitude (strength) and direction (positive versus negative) of edges can be interpreted (Epskamp et al., 2018). The most common model to estimate and visualize networks is the *partial correlation network*, i.e., edges between two nodes are estimated based on the correlation matrix, controlling for all other nodes in the network. Regularization techniques can be used to remove spurious edges, resulting in networks that are simpler to interpret (Epskamp & Fried, 2018). After estimating a network, several indices can be computed to analyse the network structure. The most frequently used indices are the centrality measures *strength, betweenness,* and *closeness,* which inform about the structural importance of individual nodes within the network. Important nodes influence other nodes in the network more strongly than less important nodes, and thus are an optimal starting point for targeted interventions (Costantini et al., 2015). Moreover, by performing a network comparison test (NCT, van Borkulo et al., 2017) networks from different education systems can be compared pair wise with regard to the invariance of their overall structure (i.e., patterning of unique interactions between indicators); invariance of their global strength (i.e., frequent co-occurrence of indicators); and invariant strength of specific edges. The advantage of network analysis in a cross-cultural context is that relations between indicators can be compared meaningfully without requiring scalar invariance (valid comparisons of correlations require metric invariance only). Thus, network analysis is a well-suited

diagnostic tool to explore whether education systems differ in their structure and dynamics of teaching and with regard to the most central and influential practice.

2.2.2 Cultural variations in the cognitive processing of survey items

Besides investigating the structure of teaching across education systems, it is equally important to understand if and how these variations in teaching as well as other cultural variations between education systems may have a bearing on how survey items are cognitively processed.

The processing of survey items is assumed to involve several interrelated cognitive stages. One prominent conceptualization is the Model of Response Process, which proposes four cognitive stages of survey responding, namely 1) item comprehension, 2) retrieval of relevant information from memory, 3) judgment formation based on the respondent's interpretation of the item in combination with the retrieved memory, and 4) response selection and response editing for reasons of consistency, acceptability, and social desirability (Tourangeau, 1984; Tourangeau et al., 2000). Errors may occur at all stages, which can threaten *cognitive validity*, i.e., the degree to which the respondents' cognitive processes mirror those intended by the researcher (Karabenick et al., 2007). Initial evidence by Lenske (2016) and Lenske and Praetorius (2020) suggests that German students often do not process teaching quality items in a valid manner. This was foremost linked to errors occurring at the *comprehension stage*. In both studies, the majority of interviewed students did not interpret items measuring classroom management and cognitive activation as intended. If students understood the items correctly, they mostly associated the item content with relevant experiences in classroom *(information retrieval stage)* and aggregated and weighed information correctly (*judgment stage).* Additionally, most students selected answering categories congruent with the information they retrieved from memory (*response stage);* yet the students' ability to select the appropriate category decreased with an increasing number of available categories (Lenske, 2016).

There is initial evidence that cultures differ in how they cognitively process survey items (e.g., Schwarz, 2007; Schwarz et al., 2010; Uskul & Oyserman, 2006; Varnum et al., 2010). In cross-cultural measurements of teaching, the involvement of various languages is probably the most serious threat

to a comparable item interpretation *(comprehension stage)*. Translation errors (Davidov et al., 2014; Fitzgerald et al., 2011) and differences regarding the complexity of translations (van de Vijver & Leung, 1997) can alter the meaning of items across cultures. Moreover, the language that is used to describe classroom phenomena often reflects culture-specific pedagogical histories and norms of practice, which can impact a comparable translation of teaching measures in two ways: First, some teaching practices may only be known in specific education systems and thus there are no adequate words to describe them in other languages. On the other hand, words describing classroom experiences may have a different connotation across education systems (Mesiti & Clarke, 2017). Hence, a culturally shaped translation may trigger culture-specific interpretations and thus compromise the comparative validity of teaching measures. Likewise, the *information retrieval and judgement stage* are subjected to cultural influences. When responding to items measuring teaching, students' have to recall classroom experiences relevant to the item content. Yet, cultures differ in how they perceive experiences, how they store and organize information in memory, and the extent to which their perceptions of experiences are influenced by internal states (Schwarz et al., 2010). Thus, when thinking of the same experience in classroom, students from different education systems may differ in their judgments of that experience. Moreover, as outlined in the previous section, teaching likely differs between education systems and experiences that students make in classrooms around the world are not necessarily the same (Stigler & Hiebert, 1999). Consequently, the recalled memory may vary between education systems, which can have a bearing on the interpretation of items and constructs. And lastly, non-invariance of items measuring teaching may be caused by culture-specific response styles, i.e., a systematic tendency to endorse certain response options on some basis other than the item content (*response selection stage,* Paulhus, 1991). Most notably, respondents from different cultures have been shown to differ in their use of rating scales and to either have a tendency to the positive side (acquiescence responding) or the negative side (disacquiescence responding), the extremes (extreme responding), or the centre of the scale (midpoint responding, Baumgartner & Weijters, 2017). Culture-specific response styles can lead to erroneous conclusions of cross-cultural

differences in teaching as survey responses vary between cultural groups even though there are no true differences in teaching - or vice versa (Baumgartner & Steenkamp, 2001).

Thus, different scores from the assessment might represent true differences in teaching, cultural differences in the response process, or an unknown combination of both. Therefore, it is crucial to unfold the role played by culture in the response process. Despite its critical relevance, so far research has paid limited attention to the impact of culture on the cognitive processing of teaching items, which is the fourth aim of this dissertation. Understanding cultural similarities and differences in the response process can provide valuable insight into potential sources of non-invariance and assist in comparable measurement.

Dissertation aim 4: *The fourth aim is to investigate sources of the limited invariance of teaching measures with regard to cultural variations in the cognitive processing of survey items (Manuscript 2)*

A method frequently used for gaining insight into the cognitive processing of survey items is ***cognitive interviewing*** (**CI,** Beatty & Willis, 2007; Koskey et al., 2010). In a cross-cultural context, CI ideally provides evidence that there are no cultural differences in the cognitive processing of survey items. Otherwise, it has the potential to identify sources of variation that occur at different cognitive stages. Therefore, verbal information is collected about how survey items are approached, consumed, and digested (Beatty & Willis, 2007; Willis & Miller, 2011).

Two major techniques for prompting respondents to verbalize their thought processes can be distinguished (Willis, 2005). *Think-aloud* (also concurrent verbalization) requires the respondents to verbalize their thoughts during each item response without being disturbed by the interviewer. Think-aloud is expected to capture actual thought processes and to hardly be affected by interviewer bias. However, respondents often have difficulties to freely think aloud (Willis, 2005) or do not provide enough relevant information (Conrad & Blair, 2004). Further, think-aloud has been criticized for not being well-suited to identify cross-group differences in survey responding as it may function differently across cultures (Pan, 2004; Willis, 2015). *Verbal probing* is characterized by an interviewer asking

follow-up questions to elicit additional information during (concurrent probing) or after the respondent has completed the questionnaire (retrospective probing) (Willis, 2005). Probes help to focus on information that is important to the researcher and are designed to further gauge the extent to which respondents interpret and answer items as intended (Desimone & Le Floch, 2004, for a compendium of different types of probes see Willis, 2005). There is a broad consensus that probing is a compelling means to evaluate the cross-cultural comparability of survey items and that it works well in a vast majority of cultures; yet the effectiveness of specific probes may vary across cultures (Willis, 2015). Further, to consider cross-cultural differences in the response processes and at the same time ensure comparable probing results, Miller and colleagues (2011) recommend combining standardized probes that are the same for all respondents with emergent probes that are matched to the interviewing situation and cultural context. It is equally important to discuss and analyse results jointly with all collaborators of a study in order to avoid a culturally biased interpretation of the interview data.

3 Empirical studies: summary of manuscripts

In the following, a summary of the research aims, methods, and results is provided for the three manuscripts that address the dissertation aims empirically. The manuscripts can be found in Appendices A to C.

3.1 Manuscript 1: Using alignment to check invariance of teaching quality measures across and within linguistic clusters

Fischer, J., Praetorius, A.-K., & Klieme, E. (2019). The impact of linguistic similarity on cross-cultural comparability of students' perceptions of teaching quality. *Educational Assessment, Evaluation and Accountability, 31,* 201-220. https://doi.org/10.1007/s11092-019-09295-7

3.1.1 Aims

Manuscript 1 investigates the cross-cultural invariance of classroom management, student support, and cognitive activation measures by using a sophisticated method, namely alignment *(Dissertation aim 1).* Additionally, Manuscript 1 investigates the impact of linguistic similarity on the degree of the cross-cultural invariance of teaching quality measures *(Dissertation aim 2)*.

3.1.2 Methods

Measures and sample. The analysis was based on PISA 2012 data of students' perceptions of classroom management, student support, and cognitive activation in mathematics instruction (OECD, 2014). To investigate the impact of linguistic similarity on measurement invariance, the sample consisted of 15 education systems (selected according to several criteria, see Appendix A) which can be grouped into five major linguistic clusters (Chinese-speaking cluster: Chinese-speaking Macao, Shanghai, Taipei; English-speaking cluster: English-speaking Ireland, England and Wales, Scotland; French-speaking-cluster: Belgium, France, French-speaking Switzerland; German-speaking cluster: Austria, Germany, German-speaking Switzerland; Spanish-speaking cluster: Chile, Colombia, Mexico). Students with missing data on all items were excluded from analysis. To avoid different model contributions due to varying sample sizes, a random subsample of 1,000 students per education system was drawn according to final student weights (3,000 students per linguistic cluster, N=15,000).

Measurement invariance testing. To check measurement invariance, the alignment method by Asparouhov & Muthén (2014) was applied (see Section 2.1.1). Analysis accounted for the hierarchical PISA data structure (students nested in schools and education systems) and the MLR estimator for parameter estimates that are robust to non-normality and non-independence of observations was applied. Since standard errors indicated a poor model fit using free alignment, the fixed estimation method was applied. The level of invariance was inferred according to the rule of thumb suggested by Flake and McCoach (2018) based on simulation studies, whereupon latent means can be compared meaningfully if less than 29% item intercepts of a scale are non-invariant. Analysis was conducted in two steps: 1) checking measurement invariance across all education systems (not controlling for linguistic similarity) separately for student support, classroom management, and cognitive activation (three models); and 2) checking measurement invariance within each linguistic cluster for every dimension (15 models, resulting in a total of 18 models).

3.1.3 Results

The main results listed in Manuscript 1 include:

1) *Factor loading* non-invariance was exceedingly low and approximate metric invariance was met in all models (between 0 and 8 % factor loading non-invariance). Thus, associations between the teaching quality dimensions and other variables can be compared across the 15 education systems.
2) The amount of non-invariant *intercepts* was relatively high, overall. In the models with all 15 education systems, intercept non-invariance was the highest for classroom management, followed by cognitive activation and student support (37, 33, and 32% non-invariant intercepts, respectively). Consequently, the degree of non-invariance was above the threshold of 29% non-invariant intercepts for all three dimensions. Thus, contrary to the assumption expressed in *Dissertation aim 1*, a more flexible method to test measurement invariance did not yield approximate scalar invariance, implying that mean differences in teaching quality cannot be compared validly across the 15 education systems.
3) Within the five linguistic clusters, intercept non-invariance was much lower (0 to 22% non-invariance). Consequently, other than across all 15 education systems, approximate scalar

invariance was met for all dimensions within the linguistic clusters. Thus, by considering linguistic similarity, means can be compared across a subset of education systems, which is in line with expectations outlined in *Dissertation aim 2*.

4) Item characteristics had an impact on measurement invariance. While national adaptions seemed to enhance the level of invariance, complex items involving more than one concept (e.g., the teacher shows an interest in students learning) and items with ambiguous wordings (e.g., extended time or complex problems) showed a comparably low cross-cultural invariance. Further, items focusing on the students' *understanding* were invariant across the 15 education systems, while the concept and metrics of items targeting the students' *learning* differed.

3.2 Manuscript 2: Understanding the lack of cross-cultural invariance of teaching quality measures - students' interpretations of student support items in Germany and China

Fischer, J., Klieme, E., Praetorius, A.-K., Jinjie, X. (submitted). Understanding lack of equivalence in cross-cultural measurements of teaching quality: Students' interpretations of student support items in Germany and China. *Submitted to Teaching and Teacher Education.*

3.2.1 Aims

Manuscript 2 investigates cultural similarities and differences in the cognitive processes of survey responding as a potential source of the limited invariance of student support items (*Dissertation aim 4).* More precisely, Manuscript 2 investigates how students from Germany and China (Shanghai) interpret student support items, and which experiences in classroom they associate with the item content (see Appendix B for the reasons for selecting Germany and China (Shanghai) and for evaluating student support items). To link qualitative differences in item interpretation to the statistical analysis of measurement invariance, Manuscript 2 additionally tests the assumptions that a) item interpretation and associations hardly differ between students from Germany and China (Shanghai) for statistically invariant student support items, while b) interpretations and associations are expected to differ more strongly for statistically non-invariant student support items.

3.2.2 Methods

Selection of items for Cognitive interviews (CI) and measurement invariance testing. To study similarities and differences in the cognitive processing of student support items, CIs were conducted with students from Germany and Shanghai (China). To ensure a reasonable response burden during CIs and to link qualitative and quantitative evidence, first two non-invariant and one invariant item were selected out of the five PISA student support items used in Manuscript 1. To this effect, alignment was applied for a random subsample of 1,000 students each from China (Shanghai) and Germany based on PISA 2012 data. The model specifications were identical to Manuscript 1.

Results.

1) All factor loadings were invariant (approximate metric invariance was reached).
2) Based on intercept non-invariance, Item TS01 (*"The teacher shows an interest in every student's learning")* and TS04 *("The teacher helps students with their learning")* were selected as the two non-invariant items for the CIs. As measurement invariance was particularly limited for items involving the term *learning* in Manuscript 1*,* the aim of Manuscript 2 was to identify potential sources of non-invariance occurring in the survey response process.
3) TS05 (*"The teacher continues teaching until students understand")* was selected as the invariant item (invariant intercepts for Germany and China, Shanghai). For invariant items, CI can provide evidence if interpretative patterns are indeed similar across education systems as suggested by the statistical analysis. As more than one item was invariant, the selection was based on theoretical considerations, such as the goal to cover wide aspects of the construct.

Evaluating item interpretation with Cognitive interviews (CI).

Sample. To evaluate the selected items, cognitive interviews were conducted with 14 native speaking students each from China (Shanghai) and Germany (N=28). As the aim was to evaluate PISA student support items as framed for mathematics instruction, the students were chosen to match the background of PISA 2012 student respondents. All selected students attended schools on the secondary level. In China (Shanghai), 57 percent of the sampled students were male (PISA 2012: 49% male), had an average last math grade of 123 (equals "good"), and an average of 80 books at home

(PISA 2012: 100 books). In Germany, 57 percent of the students were male (PISA 2012: 51% male), had an average last math grade of three (equals "satisfying"), and 175 books at home on average (PISA 2012: 170 books).

Instruments. After answering a context questionnaire, the students verbally completed a questionnaire ("think-aloud") consisting of the three selected items, worded identically as in the German/Shanghai PISA 2012 student questionnaire. Subsequently, they answered follow-up questions during a semi-structured retrospective probing. Think-aloud and probing were pretested prior to data collection. However, most students had difficulties to verbally express their thoughts in a monolog and thus the interpretative value of the think-aloud was rather low. Hence, in Manuscript 2, only the results of the probing were reported. The probing protocol consisted of a set of pre-scripted standardized probes targeting in depth information on 1) item interpretation including key terms and 2) associations of the item content with experiences in classroom. Additionally, interviewers could ask spontaneous probes matched to the cultural context and situation. To eliminate interviewer effects and to standardize the interviews as much as possible, the same (German) interviewer probed the German and Chinese (Shanghai) students (with the help of the Chinese research team and an interpreter in Shanghai). The average duration of the probing was 10 minutes in China (Shanghai) and 12 minutes in Germany. The interviews were audio-recorded, and both the German and Chinese (Shanghai) audio data were transcribed in German (according to rules defined by Kuckartz, 2018).

Qualitative content analysis. The interview data was coded and comparatively analysed with computer-aided Qualitative Content Analysis (Kuckartz, 2018). The coding system, consisting of main and sub codes, was developed in several cycles. First, one (German) coder developed a coding system based on the German data (data-driven). This coding system was revised based on the input of a second independent coder and an expert's input on teaching and learning (deductive coding). The resulting coding system was then used to re-code the German data. In a second step, one (German) coder applied the coding system to code the Chinese (Shanghai) data and codes were added or existing codes revised if necessary. The resulting coding system was revised based on feedback provided by

the second coder, the Chinese research team, and experts on qualitative research and instruction. The final coding system was then used to re-code all interviews. Afterwards, the assigned codes were comparatively analysed to identify differences and similarities in item interpretation between students from China (Shanghai) and Germany. Chi-square tests were applied to evaluate whether differences were significant, and frequencies of assigned main and sub codes were computed to aid the interpretation of findings.

3.2.3 Results

The main findings of the CIs include:

1) Regardless of group membership, students frequently associated the student support items with teaching practices supporting their competence, such as remedial activities with the aim to eliminate learning problems.
2) Besides this interpretative overlap, CIs revealed distinct interpretative variations between the interviewed students from Germany and China (Shanghai). This supports assumptions as addressed by *Dissertation aim 4,* whereupon cultures may differ in the cognitive processing of survey items. CIs identified several sources of variation:

a) Differences in the translation of the key term *learning* (in Germany *learning progress*, in Shanghai *learning state,* Item TS01) produced pronounced interpretative variations, which might explain the non-invariance found in Manuscript 1 for the same item.
b) Items targeting behaviours that are not directly observable for students (e.g., the teachers' interest, TS01) and items with ambiguous wordings (e.g., teaching *until* students understand, TS05) increased the likelihood of culture-specific interpretations. These variations resembled differences in preferred instructional approaches and goals (e.g., time management or targeted student group, see Leung, 2001). Thus, CIs also provided initial evidence that education systems differ in their designs of teaching as is addressed by *Dissertation aim 3*.
c) Moreover, CIs identified differences in the definition of supportive practices as potential source of interpretative variations. Besides competence support, German students often referred to

practices supporting socio-emotional experiences and autonomy, while Chinese students associated the items with practices supporting their academic productiveness (e.g., retaining students' attention). These, however, are no sub-dimensions of student support, but classroom management (Klieme et al., 2009).

d) While interpretative variations were more pronounced for statistically non-invariant items, CIs also identified interpretative variations for the statistically invariant item. Thus, the results of the CIs do not support the results of the statistical analysis in Manuscript 1 and 2, whereupon the item targeting students understanding (TS05) is (statistically) comparable across education systems.

3.3 Manuscript 3: Identifying cultural differences and similarities in the structure of teaching practices: combining MGCFA and network analysis

Fischer, J., He, J., & Klieme, E. (2020). The structure of teaching practices across countries: A combination of factor analysis and network analysis. *Studies in Educational Evaluation, 65.* https://doi.org/10.1016/j.stueduc.2020.100861

3.3.1 Aims

Manuscript 3 investigates similarities and differences in the structure of teaching practices across education systems as a potential source of the limited invariance of teaching practice measures in two steps (*Dissertation aim 3)*. First, Manuscript 3 investigates whether the theoretical distinction between teacher-directed and student-centred practices is empirically supported across education systems. Additionally, Manuscript 3 investigates differences and similarities in the structure and co-occurrence of teaching practices across education systems (by considering linguistic similarity) and explores how individual assessment practices relate to teacher-directed and student-centred teaching practices.

3.3.2 Methods

Measures and sample. The analysis was based on PISA 2012 data of students' perceptions of teacher-directed, student-centred, and classroom assessment practices in mathematics instruction (OECD, 2014). Given the results from Manuscript 1, whereupon linguistic similarity may enhance the degree of measurement invariance, the sample of Manuscript 3 consisted of 12 education systems grouped into four major linguistic clusters. Besides language, the clusters differed in their affluence level, power

distance, and values of individualism and collectivism, which can have a bearing on the perceptions and preferences of teaching practices (Chinese-speaking cluster: Macao, Shanghai, Taipei; English-speaking cluster: Australia, United Kingdom, United States; German-speaking cluster: Austria, Germany, Switzerland; Spanish-speaking cluster: Chile, Colombia, Mexico). To rule out method artefacts due to missing values and varying sample sizes, a random subsample of 1,000 students per education system with complete responses on the targeted items was drawn according to final student weights (N=12,000).

Measurement invariance testing with MGCFA. To test if teacher-directed and student-centred practices are two distinct factors across the 12 education systems, measurement invariance of a two-factor model was checked with MGCFA. Afterwards, to test if classroom assessment forms a third factor in the teaching practice framework (see Section 1.1), classroom assessment was added in a three-factor MGCFA across the 12 education systems. The model fit was evaluated based on recommendations by Cheung and Rensvold (2002) and Hu and Bentler (1999) (acceptable if CFI > .90, RMSEA and SRMR <.08). The level of measurement invariance was evaluated by adhering to the rule of thumb by Rutkowski and Svetina (2014) (cut point of change CFI: .02 and RMSEA: .03 from the configural to the metric model, and from the metric to the scalar model the changes of both CFI and RMSEA should be within .01).

Checking the structure of teaching practices with network analysis. In a second step, network analysis (see Section 2.2.1) was performed to explore the structure and co-occurrence of teaching practices across education systems. For each education system, a partial correlation network was estimated and visualized. A regression-based filtering approach (LASSO) was incorporated for a sparse and more interpretable model (Costantini et al., 2015). After ensuring the accuracy and stability of estimates based on a nonparametric bootstrapping test for each network (Epskamp & Fried, 2017), three sets of analysis were performed: 1) pair-wise comparisons (=66 comparisons) of the invariance of the overall network structure and global connectivity (with significant testing based on permutations, Network comparison test, van Borkulo et al., 2016); 2) within the networks a) similarities and differences in the

importance of individual teaching practices (strength-centrality) were checked and b) the relation of individual classroom assessment practices with either teacher-directed or student-centred practices was compared across education systems (bootstrapped edge difference test, Epskamp & Fried, 2017).

3.3.3 Results

Results of the **measurement invariance testing with MGCFA** include**:**

1) Teacher-directed and student-centred practices reached metric invariance [$p < .01$, CFI =.907, RMSEA =.069, SRMR =.070]. Thus, the theoretical distinction is generalizable across the 12 education systems. Yet, in line with findings of Manuscript 1 for teaching quality measures, teacher-directed and student-centred practices did not reach scalar invariance [$p < .01$, CFI =.740, RMSEA =.109, SRMR =0.089]. This may be due to intrinsic differences in metrics of these two constructs or methodological artefacts that prevented valid comparisons on mean levels across the 12 education systems.
2) After adding classroom assessment practices, configural invariance was just accepted [$p < .01$, CFI =.900, RMSEA =.076, SRMR =.050]. Thus, across the 12 education systems, classroom assessment was no third comparable factor in the teaching practice framework.

Checking the structure of teaching practices with network analysis. The nonparametric bootstrapping testing based on 1000 bootstrapped samples supported the accuracy of the networks and the strength centrality showed acceptable stability, allowing valid inferences. In general, the results emphasize culture-specific structures and patterns of teaching practices as addressed by *Dissertation aim 3*, which may explain the limited invariance of teaching practice measures found with MGCFA:

1) Across the networks, most teaching practices were positively mutually linked (even teacher-directed and student-centred practices), indicating that the more frequent application of one practice seems to go hand in hand with the more frequent application of another connected practice, conditioning on all remaining practices. Yet, the strength of the edge weights differed across education systems.

2) The *overall network structure* and to a lesser extent the *global connectivity* significantly differed for most pair wise comparisons. Thus, the education systems not only differed in their patterning of unique relations between teaching practices, but also in the extent to which teaching practices frequently co-occur. Even though this was not supported for all within-cluster pair wise comparisons, the network structure was more similar *within* compared to *across* linguistic clusters (which is in line with findings from Manuscript 1 and supports assumptions addressed by *Dissertation aim 2*).
3) The centrality of individual teaching practices within the network varied across the 12 education systems. Central practices accompany and are easily aligned with many other practices and thus are an optimal starting point for targeted interventions. Particularly the assessment practice of providing individualized feedback on strength and weaknesses played a central role in the networks of most education systems and seems to be at the heart of teaching as perceived by students. However, the centrality of the remaining practices showed a less clear patterning across education systems.
4) Instead of clustering together, the individual classroom assessment practices either more strongly related to teacher-directed or student-centred practices, which explains why classroom assessment formed no stand-alone latent factor in the MGCFA. Across all networks, assessment practices that are used to structure and guide classroom learning were significantly more strongly conditionally related to teacher-directed practices. Assessment practices supporting individualized learning showed a less common patterning across the networks; yet tended to be more strongly related to student-centred practices.

4 Discussion

As noted by He and van de Vijver (2013, p. 51): "[…] the quality of cross-cultural educational studies could be further improved by paying more attention to methodological issues". With a focus on questionnaire measures of teaching, this dissertation adopts this recommendation and aims to provide a comprehensive understanding of issues of measurement invariance and to contribute to a more comparable cross-cultural measurement of teaching. Four dissertation aims were formulated and addressed in three empirical studies. *Manuscript 1* checked the invariance of teaching quality measures across education systems using alignment. Additionally, the impact of cultural proximity (operationalized as linguistic similarity) of the compared education systems on the degree of measurement invariance was investigated. *Manuscript 2* and *3* zoomed into the impact of culture and investigated cultural variations in the cognitive processing of teaching quality items and in the structure and dynamics of teaching practices as potential sources of bias (including checks of measurement invariance with MGCFA).

In the following, the findings are discussed synoptically and implications for cross-cultural comparisons of teaching (Section 4.1.1), instrument development (Section 4.1.2), analysis (Section 4.1.3), and the conceptualization of teaching in a cross-cultural context (Section 4.1.4) are derived. Thereafter, their relevance for cross-cultural educational psychology (Section 4.2) and limitations and future directions (Section 4.3) are illustrated and finally this dissertation is rounded off by an overall conclusion (Section 5).

4.1 Discussion and implications

4.1.1 Comparing teaching across education systems

If researchers are interested in comparing associations of teaching quality and practices with other variables across multiple education systems, bias in all likelihood does not compromise the validity of interpretations, since measures of teaching quality and teacher-directed and student-centred practices reached (approximate) metric invariance (Manuscripts 1, 3). These findings coincide with previous measurement invariance investigations that have demonstrated metric invariance for

educational constructs (e.g., He & Kubacka, 2015). However, the lack of (approximate) scalar invariance across multiple education systems emphasizes that researchers should refrain from evaluating any cross-cultural comparisons of mean differences in teaching without having tested for measurement invariance (i.e., ranking education systems according to mean values without prior tests of measurement invariance as practiced by PISA 2012, OECD, 2013c). Otherwise, they risk drawing erroneous inferences, which may misinform and mislead cross-cultural educational research and educational policy decisions. When scalar invariance fails across multiple education systems, researchers should set out to investigate if there are clusters of education systems with invariant measurement for which comparisons of mean differences may nevertheless be meaningful. Findings of Manuscript 1 suggest that linguistic similarity has an impact on the degree of measurement invariance and education systems can thus be clustered accordingly. Contrary to across multiple education systems, measures of teaching quality reached (approximate) scalar invariance within five linguistic clusters, each comprising education systems with similar language (*Dissertation aim 2*). These findings extend findings from Klieme (2020) and Scherer and colleagues (2016), who demonstrated scalar invariance across English-speaking education systems. Thus, by considering linguistic similarity, mean differences in teaching may be compared meaningfully within clusters of education systems. With the novel mixture multigroup factor analysis (MMG-FA, de Roover, accepted), researchers are no longer required to cluster education systems for which scalar invariance may hold prior to data analysis. Rather, MMG-FA automatically generates clusters of groups wherein latent means can be validly compared. Moreover, it may help researchers to identify cultural factors besides linguistic similarity, which may have an impact on measurement invariance. For instance, clusters of education systems with invariant measurement may represent education systems with similar teaching practice profiles as suggested by network analysis for English-speaking education systems in Manuscript 3.

4.1.2 Instrument development

Across multiple education systems, scalar invariance could neither be established with MGCFA nor with the more flexible alignment method (*Dissertation aim 1*). Thus, the lack of scalar invariance of

teaching measures that has been demonstrated with MGCFA (e.g., He & Kubacka, 2015; Nilsen & Gustafsson, 2016) may not only be caused by unrealistically strict assumptions of equal parameters across cultural groups as hypothesized in Section 2.1.1. Rather, existing survey instruments do not seem well-suited for cross-cultural comparisons of mean differences in teaching. The following recommendations, which are based on the empirical findings with regard to sources of cultural bias (*Dissertation aims 3 and 4)*, may help to enhance the cross-cultural comparability of questionnaires:

First, cross-cultural studies should implement rigorous translation and verification procedures to ensure the equivalence between various national and linguistic questionnaires. To date, equivalence checks are mostly limited to the equivalence between the respective national version and the source questionnaire (e.g., PISA see OECD, 2014). Hence, translation variations between national versions may often remain undetected, such as the culture-specific translation of the term *learning* which led to pronounced variations in item interpretation between German and Chinese (Shanghai) students (Manuscript 2). Besides cross-checking national translations (e.g., with back-translations), cognitive interviews are a suitable means to detect translation variations and therefore should be an integral part of the instrument development process of every cross-cultural study.

Moreover, cross-cultural studies should avoid complex and ambiguous items. These items not only displayed a low cross-cultural invariance (Manuscript 1), but also led to pronounced culture-specific variations in item interpretation (Manuscript 2). Therefore, they may be unsuitable for cross-cultural comparisons. This includes a) items involving more than one concept (e.g., the teacher shows an interest in students' learning). If items are too complex, students tend to reduce complexity, which may alter the item meaning (Lenske, 2016); b) items focusing on abstract concepts instead of overt behaviors (e.g., the teacher's interest). If behaviors are unobservable for students, they have to rely on indirect indicators, which they may misinterpret (Fauth et al., 2020); and c) items with ambiguous formulations (e.g., extended time or complex problems) as these may leave ample room for culture-specific interpretations. On the other hand, shorter, simpler, and unambiguous items are expected to enhance the cross-cultural invariance of questionnaire measures (Harkness et al., 2003).

Finally, cross-cultural studies should refrain from the concept of strict equivalence between questionnaire versions of different cultural groups and rather focus on the cross-cultural suitability of survey instruments. In this dissertation, I point out the many differences between cultures that may have an impact on survey response, ranging from differences in the structures and dynamics of teaching (Manuscript 3) to differences in how information is cognitively processed (Manuscript 2). Thus, to ensure the cross-cultural suitability of questionnaires, it is important to acknowledge these cultural variations instead of administering identical instruments across various cultures which have been developed based on theories and models for a certain context that may not be transferable to others. One means of enhancing the cross-cultural suitability of questionnaires is by national adaptions, which usually amount to a combination of a literal translation of some stimuli and a change of other stimuli when a close translation would be inadequate for linguistic, cultural, or psychometric reasons (He & van de Vijver, 2012). The findings favor the utilization of national adaptions: the only item that required a national adaption showed the lowest amount of non-invariant item intercepts (Manuscript 1). Furthermore, cognitive interviews (Manuscript 2) provide examples of how national adaptions may enhance the cross-cultural suitability of student support items. For instance, cognitive interviews revealed that in China (Shanghai) supportive practices foremost take place after the actual instruction. The evaluated PISA student support items, however, specifically refer to the actual lessons and thus student support may be underestimated in China (Shanghai). This may be mitigated by adapting the time-related reference for China (Shanghai), while the actual item content remains unchanged. Other than by adaptations, the cross-cultural suitability of questionnaire measures may be enhanced by assessing a construct with a common core of items complemented by culture-specific items (etic and emic approach, Cheung et al., 2011). For instance, cognitive interviews revealed that students from Germany and Shanghai frequently associated the student support items with practices that support their competence. In addition, student support was perceived as supporting students' academic productiveness in Shanghai (e.g., retaining students' attention), yet supporting socio-emotional experiences and autonomy in Germany. Thus, to adequately measure constructs (here

student support) across cultures, both culture-specific indicators and indicators that target universal aspects of that construct may be needed. Yet, national adaptions and culture-specific items have to be implemented carefully, as they should account for cultural variations while at the same time a certain level of comparability has to be maintained.

4.1.3 Analysis

The following recommendations regarding the choice of analysis method and procedure may additionally help to ensure valid conclusions based on cross-cultural educational data:

First, it is important to select the most appropriate psychometric method for a measurement invariance investigation. Given that the linguistic and cultural heterogeneity poses major difficulties for scalar invariant measures across dozens of education systems (Manuscript 1, 3), methods that can flexibly account for and tolerate trivial non-invariance ought to be preferred to more rigorous methods such as MGCFA. Other than alignment, various methods that relax the standard without losing the standards in invariance testing have been developed (He et al., accepted, see also Table 1). These methods are expected to better represent substantive theories, yet their extensive application in the field of cross-cultural educational psychology is still pending. To select the optimal method, Kim and colleagues (2017) suggest considering various criteria, ranging from the level of invariance needed to answer the research questions to the number of groups under investigation.

Further, the current practice of relying on statistical tests of measurement invariance may be insufficient to fully understand issues of cross-cultural invariance. By combining quantitative and qualitative evidence, this dissertation not only identified non-invariant items (quantitative evidence), but also offered potential explanations for why measurement invariance is not given (qualitative evidence). Thus, this dissertation provided nuanced information on how to enhance the cross-cultural invariance of questionnaire measures on teaching. The fact that the methods applied disagreed with regard to specific items, even adds to the importance of a mixed-methods approach. Statistically, the item targeting students understanding (TS05 of the PISA student support scale) was comparable across 15 educations systems (alignment, Manuscript 1) as well as for China (Shanghai) and Germany

(MGCFA, Manuscript 2). However, besides some interpretative overlap, cognitive interviews revealed substantive differences in item interpretation between German and Chinese (Shanghai) students. Thus, without complementing statistical tests with cognitive interviews, the presence of item bias would have remained undetected. The limited ability of most psychometric methods to identify sources of item bias may be overcome by moderated nonlinear factor analysis (MNLFA, Bauer, 2017). MNLFA allows for a full and simultaneous assessment of measurement invariance and differential item functioning. Based on CFA, group-specific covariates can be introduced to the measurement invariance model to explain and account for DIF due to cultural characteristics; yet unfortunately it has not yet been extended to multilevel and large-scale data.

And finally, the absence of scalar invariance should not discourage researchers from analyzing cross-cultural differences in teaching and utilizing the wealth of cross-cultural educational data. It is important to recognize that there are different levels of invariance, and not all inferences are contingent on fully comparable scales (He et al., accepted). Moreover, methods such as network analysis help researchers to visualize and analyze cross-cultural differences in teaching without requiring fully comparable scales. Still, measurement bias due to translation errors or cultural variations in item interpretation can nevertheless threaten the validity of analysis results based on network analysis.

4.1.4 Conceptualization of teaching in a cross-cultural context

Finally, this dissertation provides initial insights regarding the cross-cultural generalizability of teaching. Besides some commonalities, such as the prevalence of teacher-directed practices in mathematics (Manuscript 3), the findings indicate that teaching is embedded in a cultural context and therefore may only be generalizable to a limited extent. Education systems do not only seem to differ with regard to patterns and structures of teaching (i.e., the most central practice and co-occurrence of practices, Manuscript 3), but also with regard to the targeted outcomes and targeted student groups of instruction (Manuscript 2). These differences are one conceivable explanation for the limited invariance of teaching quality and practice measures, which have mostly been inspired by theoretical

frameworks based on teaching in western education systems. Theories of teaching that have been developed for a specific cultural context may be too narrow to adequately capture cultural variations in teaching and thus they are unsuitable as theoretical frameworks for cross-cultural studies. Rather, the foundation of every cross-cultural educational study should be an internationally shared conceptualisation, otherwise: "Lacking good theories […] much of what passes as cross-national comparison will be based on hunch, myth, and uninformed secondary data analysis, rather than carefully crafted national theories of education." (Rowan, 2002, p.345). To date, such an internationally shared conceptualisation that balances the need for validity within each education system and comparability across education systems does not exist for teaching quality and practices (Praetorius et al., 2019). Besides developing a joint theory based on existing conceptualizations, models, and literature from various cultural contexts (Praetorius et al., 2019), videotaped lessons from various cultures can "shine direct light on practice" (Paine et al., 2016, p.732) and help researchers to develop a joint conceptualization of teaching.

4.2 Relevance of this research for cross-cultural educational psychology

Following the discussion of findings, the following section highlights their relevance for the field of cross-cultural educational psychology.

4.2.1 Raising attention for the importance of measurement invariance testing

By discussing measurement invariance, this dissertation intended to deliver a key contribution to ensuring that testing for measurement invariance becomes an integral part of any comparative study in the future. Despite first positive developments, cross-cultural educational research has a long way ahead: while some ILSAs have incorporated formal checks of measurement invariance in their latest technical reports (e.g., PISA 2018: OECD 2019, 2020; TALIS 2018: OECD, 2019; GTI: Mihaly et al., 2021), routine psychometric checks of "cross-country validity" of the majority of ILSAs continue to involve only internal consistencies (high values indicate "comparability") and comparisons of correlations between education systems (consistent correlations are indicative of "comparability", He et al., accepted). These, however, do not provide information on the level of cross-cultural invariance.

Moreover, it is surprising to see how rarely measurement invariance testing is conducted in research practice, which may lead to cross-cultural comparisons that are inherently meaningless (Boer et al., 2018). Hence, this dissertation addresses the urgent need to keep raising the awareness of the issue of cross-cultural comparability and the importance of invariance testing to the research community (He et al., accepted).

4.2.2 Identification of sources of cultural bias

Besides drawing attention to the topic, this dissertations' most noteworthy contribution is the comprehensive investigation of issues of measurement invariance for two components of teaching, which is unparalleled to any other research in the field. While some studies have checked the cross-cultural invariance of questionnaire measures of educational constructs (e.g., Lafontaine et al., 2019; Scherer et al., 2016), no effort has yet been made to understand why those measures often work differently across cultures. By focusing on two potential sources of cultural variations, this dissertation identified a complex interaction between instrument characteristics, cultural variations in the cognitive processing of survey items, and variations in teaching across education systems as potential sources of cultural bias. These findings can guide a comparable measurement of teaching in the future.

4.2.3 Identification of strategies for dealing with scalar non-invariance

Furthermore, this dissertation identified strategies for dealing with the often demonstrated scalar non-invariance of questionnaire measures of educational constructs. Both strategies account for the many cultural and linguistic differences of education systems, either in form of flexible model assumptions of psychometric methods or by clustering education systems with invariant measurement. Consequently, this dissertation illustrates that the wealth of data on educational constructs (i.e., from ILSAs) can be utilized for mean comparisons after all, at least for a subset of education systems with invariant measurement.

4.2.4 Applying the rich toolbox of methods

Finally, this dissertation emphasizes the importance of using a variety of methods as well as the importance of choosing the optimal method of analysis. A sophisticated method was applied to check

measurement invariance that incorporates statistical rigor with flexibility and thus holds promises for maximizing the research and policy potential of cross-cultural educational data (Boer et al., 2018). Therefore, a contribution is made to the advancement of measurement invariance testing in the field of cross-cultural educational psychology, which to date has predominantly applied the often criticized MGCFA. Moreover, the substantial advantages of matching the analysis method to the research aims and the characteristics of the targeted construct is showcased. Besides introducing network analysis to the field, the theoretical assumption of a dynamic nature of teaching was considered during analysis, and first critical explorative insights were provided with regard to similarities and differences in the structures of teaching practices across education systems. Finally and probably most importantly, methods with unique strengths and weaknesses were thoughtfully combined and thus a deeper understanding of invariance and sources of cultural bias is provided. The advantages of a mixed-methods approach in a cross-cultural context have been highlighted and demonstrated in the literature (e.g., Benítez & Padilla, 2014; Benítez et al., 2019; Meitinger, 2017). However, this dissertation is the first study that has combined various quantitative and qualitative methods to study issues related to the cross-cultural invariance of teaching measures. By complementing alignment with cognitive interviews and MGCFA with network analysis, items, constructs, and education systems for which cross-cultural comparisons of teaching seem to be problematic were identified and sources of cultural variations were determined that may explain the limited invariance. Hence, instead of simply removing non-invariant items (which may cause construct underrepresentation, He et al., accepted), constructs, or education systems from comparisons (which may cause a loss of information), the mixed-methods approach provides guidance and starting points on how to enhance the cross-cultural invariance of questionnaire measures on teaching.

4.3 Limitations and future directions

The following section describes limitations of this research that may have an impact on how these findings can be utilized for cross-cultural educational psychology. Building on these limitations, directions for future research are derived.

4.3.1 Generalizability of findings

As PISA student questionnaire data are frequently used in the field of cross-cultural educational psychology, this dissertation focuses on issues of measurement invariance with regard to PISA measures of students' perceptions of teaching quality and practices. Yet, bias is always a function of an instrument applied in a specific context (Benítez et al., 2019), and further research should investigate if these findings are generalizable for additional a) *perspectives*, i.e., teachers' interpretations and responses to survey items may deviate from those of students (Fauth et al., 2020) and show different patterns of cultural variation; b) *constructs*, such as opportunity to learn (OTL). OTL is a powerful determinant of achievement (Scheerens, 2017), yet differences in curriculum may result in variations of OTL across education systems (Kuger, 2016), which may compromise the invariance of measures; c) *instruments*, which - in contrast to PISA - allow considering aspects of teaching on the classroom level (Lüdtke et al., 2009); and d) *education systems* and *broader cultural clusters* for which mean comparisons may be possible (e.g., West European, Latin American, or Asian clusters).

4.3.2 Investigation of sources of cultural bias

Cultural variations were investigated in the cognitive processing of survey items for teaching *quality* measures with cognitive interviews and cultural variations in the structures of teaching for teaching *practice* measures with network analysis. It was thus possible to study various sources of cultural bias on the construct and item level for two components of teaching, yet there are some issues that should be addressed by further research:

First, there are many more sources of cultural bias that may compromise the invariance of questionnaire measures of educational constructs. For instance, empirical investigations of how differences in administration modes may affect the cross-cultural comparability of questionnaire data are scarce in the field of cross-cultural educational psychology (e.g., paper-based versus computer-based assessment, He et al., accepted). In order to understand non-invariance and to enhance comparability, a comprehensive investigation of various sources of bias is inevitable. Secondly, cognitive interviewing has proven to be a valuable means to identify sources of cultural bias for teaching quality items, thus it is highly recommended that cognitive interviews are additionally

conducted to understand the limited invariance of teaching practice items. Network analysis enabled unique explorative insights into variations in dynamics and structures of teaching practices across education systems. However, to fully understand how these variations may compromise the invariance of teaching practice measures and to pinpoint which adaptions may be necessary to enhance their cross-cultural invariance, it is imperative to complement network analysis with cognitive interviews. It is important to note here that the findings from the cognitive interviews are limited to Chinese (Shanghai) and German students and student support items. Western and East-Asian respondents tend to show pronounced differences in basic cognitive processes (Schwarz et al., 2010) and student support is a construct that is likely to be shaped by cultural and personal perceptions. Thus, findings may vary for additional constructs and groups, which should be investigated by further research. Finally, initial findings regarding the generalizability of teaching demonstrate an urgent need to develop an internationally shared conceptualisation, which is broad enough to accommodate related, but somewhat variable teaching practices across cultures (Praetorius et al., 2019) and therefore can guide a valid and comparable measurement. To achieve this goal, the findings of this dissertation - which focused on issues of measurement invariance rather than theory development - have to be complemented by a more in-depth investigation of cultural variations in structures and dynamics of teaching, for instance with the help of video-studies.

4.3.3 Methods

This dissertation presents promising method advancements that overcome limitations of traditional methods. Admittedly, alignment and network analysis are new methods and additionally to a few existing simulation studies further research is needed to validate model specifications (e.g., the amount of non-invariance allowed with alignment), to ensure trustworthy results for various applications (e.g., in the large-scale context), and to overcome current limitations (e.g., alignment provides no model fit information). Recently, new developments have been proposed both for alignment and network analysis. Marsh and colleagues (2018) introduced an alignment-within-CFA (AwC) approach that transforms alignment from a largely exploratory tool into a confirmatory tool and

enables analyses that previously have not been possible with alignment (e.g., testing the invariance of uniquenesses and factor variances/covariances). Likewise, there is a new development for network analysis towards a better integration with classic psychometrics (Epskamp et al., 2017), and innovative research questions can be answered with information gathered in network analysis (e.g., what combination or dynamics of teaching practices especially contribute to student learning). The field of cross-cultural psychology is developing at a fast rate and it is uncertain which psychometric methods will prevail in the field. Hence, it is imperative to watch future developments and to further the validation and extension of promising innovative psychometric methods.

5 Conclusion

The valid measurement of teaching across education systems is a major challenge faced by cross-cultural educational psychology, as questionnaire measures often work differently across cultures. This dissertation demonstrates that the lack of comparability is foremost due to the great influence exerted by culture on survey response, ranging from the embeddedness of teaching in a cultural context to the impact of culture on the cognitive processing of survey items. By combining quantitative and qualitative evidence, this dissertation identified strategies for recognising and accounting for these cultural variations and therefore this dissertation considerably contributes to a more comparable cross-cultural measurement of teaching.

References

Abacioglu, C.S., Isvoranu, A.-M., Verkuyten, M., Thijs, J., & Epskamp, S. (2019). Exploring multicultural classroom dynamics: A network analysis. *Journal of School Psychology, 74,* 90-105. https://doi.org/10.1016/j.jsp.2019.02.003

Alavi, S.B., & McCormick, J. (2004). A cross-cultural analysis of the effectiveness of the Learning Organization Model in school contexts. *International Journal of Educational Management, 18,* 408–416. https://doi.org/10.1108/09513540410563112

Alexander, R.J. (2000). *Culture and pedagogy: International comparisons in primary education.* Blackwell.

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 16*, 397-438. https://doi.org/10.1080/10705510903008204

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal, 21,* 495–508. https://doi.org/10.1080/10705511.2014.919210

Bauer, D.J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods, 22*, 507–526. https://doi.org/10.1037/met0000077

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. http://dx.doi.org/10.3102/0002831209345157

Baumgartner, H., & Steenkamp, J.-B.E.M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38,* 143-156. https://doi.org/10.1509/jmkr.38.2.143.18840

Baumgartner, H., & Weijters, B. (2017). Methodological issues in cross-cultural research. In H. van Herk & C. Torelli (Eds.), *Cross cultural issues in consumer science and consumer psychology.* Springer. https://doi.org/10.1007/978-3-319-65091-3_10

Beatty, P.C., & Willis, G.B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly, 71,* 287–311. https://doi.org/10.1093/poq/nfm006

Benítez, I., & Padilla, J.-L. (2014). Analysis of nonequivalent assessments across different linguistic groups using a mixed methods approach: Understanding the causes of differential item functioning by cognitive interviewing. *Journal of Mixed Methods Research, 8,* 52-68. https://doi.org/10.1177/1558689813488245

Benítez, I., van de Vijver, F.J.R., & Padilla, J.-L. (2019). A mixed methods approach to the analysis of bias in cross-cultural studies. *Sociological Methods & Research*, *8*. https://doi.org/10.1177/0049124119852390

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*. https://doi.org/10.1007/s11092-008-9068-5

Boehnke, K., Lietz, P., Schreier, M., & Wilhelm, A.W. (2010). Sampling: The selection of cases for culturally comparative psychological research. In D. Matsumoto & F.J.R. van de Vijver (Eds.), *Cross-cultural research methods in psychology.* Cambridge University Press. https://doi.org/10.1017/CBO9780511779381.007

Boer, D., Hanke, K., & He, J. (2018). On detecting systematic measurement error in cross-cultural research: A review and critical reflection on equivalence and invariance tests. *Journal of Cross-Cultural Psychology, 49*, 713-734. https://doi.org/10.1177/0022022117749042

Borsboom, D., & Cramer, A.O.J. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology, 9,* 91–121. https://doi.org/10.1146/annurev-clinpsy-050212-185608

Byrne, B.M., Shavelson, R.J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105,* 456–466. https://doi.org/10.1037/0033-2909.105.3.456

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. https://doi.org/10.7334/psicothema2017.178

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*. https://doi.org/10.1016/j.stueduc.2016.03.005

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

Charalambous, C.Y., & Praetorius, A.-K. (2018). Studying mathematics instruction through different lenses: Setting the ground for understanding instructional quality more comprehensively. *ZDM, 50,* 355-366. https:/doi.org/10.1007/s11858-018-0914-8

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. https://doi.org/10.1207/s15328007sem0902_5

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. *American Psychologist, 66,* 593–603. https://doi.org/10.1037/a0022389

Clarke, D., & Xu, L.H. (2008). Distinguishing between mathematics classrooms in Australia, China, Japan, Korea and the USA through the lens of the distribution of responsibility for knowledge generation: Public oral interactivity and mathematical orality. *ZDM, 40,* 963-972. https://doi.org/10.1007/s11858-008-0129-5

Conrad, F.G., & Blair, J. (2004). Data quality in cognitive interviews: The case for verbal reports. In S. Presser, J.M. Rothgeb, M.P. Couper, J.L. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires.* John Wiley.

Costantini, G., Epskamp, S., Borsboom, D., Perugini, M., Mõttus, R., Waldorp, L.J., & Cramer, A.O.J. (2015). State of the aRt personality research: A tutorial on network analysis of personality data in R. *Journal of Research in Personality, 54,* 13–29. https://doi.org/10.1016/j.jrp.2014.07.003

Dalege, J., Borsboom, D., van Harreveld, F., & van der Maas, H.L.J. (2018). A network perspective on attitude strength: Testing the connectivity hypothesis. *Social Psychological and Personality Science, 10,* 746–756. https://doi.org/10.1177/1948550618781062

Davidov, E., Dülmer, H., Cieciuch, J., Kuntz, A., Seddig, D., & Schmidt, P. (2018a). Explaining measurement nonequivalence using multilevel structural equation modeling: The case of attitudes toward citizenship rights. *Sociological Methods & Research, 47,* 729–760. https://doi.org/10.1177/0049124116672678

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology, 40*, 55-75. https://doi.org/10.1146/annurev-soc-071913-043137

Davidov, E., Muthén, B., & Schmidt, P. (2018b). Measurement invariance in cross-national studies: Challenging traditional approaches and evaluating new ones. *Sociological Methods & Research, 47*, 631-636. https://doi.org/10.1177/0049124118789708

Davidov, E., Schmidt, P., & Schwartz, S.H. (2008). Bringing values back in: The adequacy of the

European Social Survey to measure values in 20 countries. *Public Opinion Quarterly, 72,* 420-45. https://doi.org/10.1093/poq/nfn035

de Roover, K. (accepted). Finding clusters of groups with measurement invariance: Unraveling intercept non-invariance with mixture multigroup factor analysis. *Structural Equation Modeling.* https://doi.org/10.31234/osf.io/5yr68

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.* OECD Publishing. https://doi.org/10.1787/5jz2kbbvlb7k-en

Desimone, L.M., & Le Floch, K.C. (2004). Are we asking the right questions? Using cognitive interviews to improve surveys in education research. *Educational Evaluation and Policy Analysis, 26,* 1-22. https://doi.org/10.3102/01623737026001001

Dewey, J. (1929). *My pedagogic creed*. Progressive Education Association.

Echazarra, A., Salinas, D., Méndez, I., Denis, V., & Rech, G. (2016). How teachers teach and students learn: Successful strategies for school. *OECD Education Working Papers*. OECD Publishing. https://doi.org/10.1787/5jm29kpt0xxx-en

Epskamp, S., & Fried, E.I. (2017). *Bootnet: Bootstrap methods for various network estimation routines*. Retrieved from https://cran.r-project.org/web/packages/bootnet/index.html

Epskamp, S., & Fried, E.I. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods, 23,* 617–634. https://doi.org/10.1037/met0000167

Epskamp, S., Maris, G., Waldrop, L.J., & Borsboom, D. (2018). Network psychometrics. In P. Irwing, T. Booth, & D.J. Hughes (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development*. Wiley. https://doi.org/10.1002/9781118489772.ch30

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized network psychometrics: Combining network and latent variable models. *Psychometrika, 82*, 904–927. https://doi.org/10.1007/s11336-017-9557-x

Fauth, B., Göllner, R., Lenske, G., Praetorius, A.-K., & Wagner, W. (2020). Who sees what? Conceptual considerations on the measurement of teaching quality from different perspectives. *Zeitschrift für Pädagogik, 66, 138-155.*

Fischer, R., & Karl, J.A. (2019). A primer to (cross-cultural) multi-group invariance testing possibilities in R. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.01507

Fitzgerald, R., Widdop, S., Gray, M., & Collins, D. (2011). Identifying sources of error in cross-national questionnaires: Application of an error source typology to cognitive interview data. *Journal of Official Statistics, 27,* 569-599.

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. https://doi.org/10.1080/10705511.2017.1374187

Fuller, B., & Clarke, P. (1994). Raising school effects while ignoring culture? Local conditions and the influence of classroom tools, rules, and pedagogy. *Review of Educational Research, 64,* 119–157. https://doi.org/10.3102/00346543064001119

Givvin, K.B., Hiebert, J., Jacobs, J.K., Hollingsworth, H., & Gallimore, R. (2005). Are there national patterns of teaching? Evidence from the TIMSS 1999 video study. *Comparative Education Review, 49.* https://doi.org/10.1086/430260

Harkness, J.A., Mohler, P.P., & van de Vijver, F.J.R. (2003). *Cross-Cultural Survey Methods*. John Wiley & Sons.

Harlen, W., & Deakin-Crick, R. (2002). *A systematic review of the impact of summative assessment*

and tests on students' motivation for learning. EPPI-Centre.

Hattie, J. (2009). *Visible learning*. Routledge.

He, J., Barrera-Pedemonte, F., & Buchholz, J. (2019). Cross-cultural comparability of noncognitive constructs in TIMSS and PISA. *Assessment in Education: Principles, Policy & Practice, 26*, 369-385. https://doi.org/10.1080/0969594X.2018.1469467

He, J., Buchholz, J., & Fischer, J. (accepted). Cross-cultural comparability of latent constructs in ILSAs. In T. Nilsen, A. Stancel-Piątak, & J.-E. Gustafsson (Eds.), *Springer international handbooks of education. International handbook of comparative large-scale studies in education: Perspectives, methods and findings.* Springer.

He, J., & Kubacka, K. (2015). Data comparability in the teaching and learning international survey (TALIS) 2008 and 2013. *OECD Education Working Papers.* OECD Publishing. https://doi.org/10.1787/5jrp6fwtmhf2-en

He, J., & van de Vijver, F.J.R. (2012). Bias and equivalence in cross-cultural research. *Online Readings in Psychology and Culture, 2*. https://doi.org/10.9707/2307-0919.1111

He, J., & van de Vijver, F.J.R. (2013). Methodological issues in cross-cultural studies in educational psychology. In G.A.D. Liem & A.B.I. Bernardo (Eds.), *Advancing cross-cultural perspectives on educational psychology. A festschrift for D.M. McInerney.* Age publishing.

Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations.* SAGE.

Hu, L.-t., & 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. https://doi.org/10.1080/10705519909540118

Hwang, H.C., & Matsumoto, D. (2013). Culture and educational psychology. In G.A.D. Liem & A.B.I. Bernardo (Eds.), *Advancing cross-cultural perspectives on educational psychology. A festschrift for D.M. McInerney.* Age publishing.

Jöreskog, K.G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36,* 409-426. https://doi.org/10.1007/BF02291366

Karabenick, S.A., Woolley, M.E., Friedel, J.M., Ammon, B.V., Blazevski, J., Bonney, C.R., de Groot, E., Gilbert, M.C., Musu, L., Kempler, T.M., & Kelly, K.L. (2007). Cognitive processing of self-report items in educational research: Do they think what we mean? *Educational Psychologist, 42,* 139–151. https://doi.org/10.1080/00461520701416231

Khorramdel, L., Pokropek, A., & van Rijn, P.W. (2020). Establishing comparability and measurement invariance in large-scale assessments, part I. *Psychological Test and Assessment Modeling, 62,* 3-10.

Kim, E.S., Cao, C., Wang, Y., & Nguyen, D.T. (2017). Measurement invariance testing with many groups: A comparison of five approaches. *Structural Equation Modeling*, *24*, 524–544. https://doi.org/10.1080/10705511.2017.1304822

Klieme E. (2020). Policies and practices of assessment: A showcase for the use (and misuse) of international large scale assessments in educational effectiveness research. In J. Hall, A. Lindorff, & P. Sammons (Eds.), *International perspectives in educational effectiveness research.* Springer. https://doi.org/10.1007/978-3-030-44810-3_7

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. Seider (Eds.), *The power of video studies in investigating teaching and learning in the classroom.* Waxman.

Klieme, E., Steinert, B., & Hochweber, J. (2010). Zur Bedeutung der Schulqualität für Unterricht und

Lernergebnisse [The impact of school quality on instruction and student outcomes]. In W. Bos, E. Klieme, & O. Köller (Eds.), *Schulische Lerngelegenheiten und Kompetenzentwicklung [Learning opportunities in school and competence development]*. *Festschrift für Jürgen Baumert.* Waxmann.

Koskey, K.L.K., Karabenick, S.A., Woolley, M.E., Bonney, C.R., & Dever, B.V. (2010). Cognitive validity of students' self-reports of classroom mastery goal structure: What students are thinking and why it matters. *Contemporary Educational Psychology, 35,* 254-263. https://doi.org/10.1016/j.cedpsych.2010.05.004

Kramsch, C.J. (1998). *Language and culture.* Oxford University Press.

Kuckartz, U. (2018). *Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung [Qualitative content analysis: Methods, practices, computer-based analysis].* Beltz Juventa.

Kuger, S. (2016). Curriculum and learning time in international school achievement studies. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Assessing contexts of learning. Methodology of educational measurement and assessment.* Springer. https://doi.org/10.1007/978-3-319-45357-6_16

Kuger, S., Klieme, E., Lüdtke, O., Schiepe-Tiska, A., & Reiss, K. (2017). Mathematikunterricht und Schülerleistung in der Sekundarstufe: Zur Validität von Schülerbefragungen in Schulleistungsstudien [Student learning in secondary school mathematics classrooms: On the validity of student reports in international large-scale studies]. *Zeitschrift für Erziehungswissenschaft, 20,* 61–98. https://doi.org/10.1007/s11618-017-0750-6

Kunter, M., & Ewald, S. (2016). Bedingungen und Effekte von Unterricht: Aktuelle Forschungsperspektiven aus der pädagogischen Psychologie [Requirements and effects of high quality teaching: current research perspectives]. In N. McElvany, W. Bos, H.G. Holtappels, M.M. Gebauer, & F. Schwabe (Eds.), *Bedingungen und Effekte guten Unterrichts* [Requirements and effects of high quality teaching]. Waxmann.

Kyllonen, P.C., & Bertling, J.P. (2014). Innovative questionnaire assessment methods to increase cross-country comparability. In L. Rutkowski, M. von Davier, & D. Rutkowski, (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*. CRC Press.

Kyriakides, L., Creemers, B., & Panayiotou, A. (2020). Developing and testing theories of educational effectiveness addressing the dynamic nature of education. In J. Hall, A. Lindorff, P. Sammons (Eds.), *International perspectives in educational effectiveness research.* Springer. https://doi.org/10.1007/978-3-030-44810-3_3

Lafontaine, D., Dupont, V., Jaegers, D., & Schillings, P. (2019). Self-concept in reading: Factor structure, cross-cultural invariance and relationships with reading achievement in an international context (PIRLS 2011). *Studies in Educational Evaluation, 60, 78-89.* https://doi.org/10.1016/j.stueduc.2018.11.005

Lenske, G. (2016). *Schülerfeedback in der Grundschule [Student feedback in elementary school].* Waxmann.

Lenske, G., & Praetorius, A.-K. (2020). Schülerfeedback – was steckt hinter dem Kreuz auf dem Fragebogen [Student feedback - what is the basis for choosing an answer on a questionnaire?] *Empirische Pädagogik, 34*, 11–29.

Leung, F.K.S. (2001). In search of an East Asian identity in mathematics education. *Educational Studies in Mathematics*, *47*, 35–51. https://doi.org/10.1023/A:1017936429620

Leung, F.K.S. (2005). Some characteristics of East Asian mathematics classrooms based on data from the TIMSS 1999 Video Study. *Educational Studies in Mathematics, 60,*199-215.

https://doi.org/10.1007/s10649-005-3835-8

Little, T.D. (2013). *Longitudinal Structural Equation Modeling.* Guilford Press.

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 modeling. *Contemporary Educational Psychology, 34,* 120–131. https://doi.org/10.1016/j.cedpsych.2008.12.001

Marsh, H.W., Guo, J., Parker, P.D., Nagengast, B., Asparouhov, T., Muthén, B., & Dicke, T. (2018). What to do when scalar invariance fails: The extended alignment method for multi-group factor analysis comparison of latent means across many groups. *Psychological Methods, 23*, 524–545. https://doi.org/10.1037/met0000113

Marsh, H.W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A.J.S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling: A Multidisciplinary Journal, 16,* 439-476. https://doi.org/10.1080/10705510903008220

Maulana, R., André, S., Helms-Lorenz, M., Ko, J., Chun, S., Shahzad, A., Irnidayanti, Y., Lee, O., de Jager, T., Coetzee, T., & Fadhilah, N. (2021). Observed teaching behaviour in secondary education across six countries: Measurement invariance and indication of cross-national variations. *School Effectiveness and School Improvement, 32.* https://doi.org/10.1080/09243453.2020.1777170

Meitinger, K. (2017). Necessary but insufficient: Why measurement invariance tests need online probing as a complementary tool. *Public Opinion Quarterly, 81,* 447-472. https://doi.org/10.1093/poq/nfx009

Mesiti, C., & Clarke, D. (2017). The international lexicon project: Giving a name to what we do. In R. Seah, M. Horne, J. Ocean, & C. Orellana (Eds.), *Proceedings of the Mathematical Association of Victoria annual conference.*

Mihaly, K., Klieme, E., Fischer, J., & Doan, S. (2021). Questionnaire scale characteristics. *In OECD (Eds.) Global Teaching InSights Technical Report.* OECD Publishing.

Miller, K., Fitzgerald, R., Padilla, J.-L., Willson, S., Widdop, S., Caspar, R., Dimov, M., Gray, M., Nunes, C., Prüfer, P., Schöbi, N., Schoua-Glusberg, A., & Willis, G.B. (2011). Design and analysis of cognitive interviews for comparative multinational testing. *Field Methods, 23,* 379–396. https://doi.org/10.1177/1525822X11414802

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods, 17*, 313–335. https://doi.org/10.1037/a0026802

Muthén, B., & Asparouhov, T. (2014). IRT studies of many groups: The alignment method. *Frontiers in Psychology, 5*. https://doi.org/10.3389/fpsyg.2014.00978

Muthén, B., & Asparouhov, T. (2018). Recent methods for the study of measurement invariance with many groups: Alignment and random effects. *Sociological Methods & Research, 47*, 637–664. https://doi.org/10.1177/0049124117701488

Nilsen, T., & Gustafsson, J.-E. (2016). *Teacher quality, instructional quality and student outcomes. Relationships across countries, cohorts and time. IEA research for education*. Springer.

Oberski, D.L. (2014). Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. *Political Analysis*, *22,* 45-60. https://doi.org/10.1093/pan/mpt014

OECD. (2013a). *PISA 2012 Assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy.* OECD Publishing. http://dx.doi.org/10.1787/9789264190511-en

OECD. (2013b). *Teaching and learning international survey TALIS 2013: Conceptual framework.* OECD Publishing.

OECD. (2013c). *PISA 2012 results: Ready to learn– students' engagement, drive and self-beliefs (volume III).* OECD Publishing.

OECD. (2014). *PISA 2012 technical report.* OECD Publishing.

OECD. (2019). *Teaching and learning international survey TALIS 2018: Technical report*. OECD Publishing.

OECD. (2019, 2020). *PISA 2018 technical report*. OECD Publishing.

Opfer, V.D., Bell, C.A., Klieme, E., McCaffrey, D.F., Schweig, J.D., & Stecher, B.M. (2020). Understanding and measuring mathematics teaching practice in eight countries and economies from four continents. In OECD (Eds.), *Global Teaching InSights: A video study of teaching*. OECD Publishing. https://doi.org/10.1787/20d6f36b-en

Osborn, M.J., Broadfoot, P.M., McNess, M.E., Ravn, B., Planel, C.D., & Triggs P.A. (2003). *A world of difference? Comparing learners across Europe.* Open University Press.

Paine, L., Blömeke, S., & Aydarova, O. (2016). Teachers and teaching in the context of globalization. In D.H. Gitomer & C.A. Bell (Eds.), *Handbook of research on teaching.* American Educational Research Association. https://doi.org/10.3102/978-0-935302-48-6_11

Pan, Y. (2004). Cognitive interviews in languages other than English: Methodological and research issues. In *Proceedings of the Survey Research Methods Section of the American Statistical Association.*

Paulhus, D.L. (1991). Measurement and control of response biases. In J.P. Robinson, P.R. Shaver, & L.S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes.* Academic Press.

Piaget, J. (1952). *The origins of intelligence in children*. International University Press.

Pianta, R.C., Hamre, B.K., & Allen, J.P. (2012). Teacher-student relationships and engagement: Conceptualizing, measuring, and improving the capacity of classroom interactions. In: S. Christenson, A. Reschly, & C. Wylie (Eds.), *Handbook of Research on Student Engagement.* Springer. http://dx.doi.org/10.1007/978-1-4614-2018-7_17

Praetorius, A.-K., Klieme, E., Bell, C.A., Qi, Y., Witherspoon, W., & Opfer, V.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.

Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018b). Generic dimensions of teaching quality: The German framework of Three Basic Dimensions. *ZDM, 50*, 407–426. https://doi.org/10.1007/s11858-018-0918-4

Praetorius, A.-K.**,** Rogh, W., Klieme, E., & Bell, C.A. (2019). Methodological challenges in conducting international research on teaching quality using standardized observations. In L.E. Suter, E. Smith, & B.D. Denman (Eds.), *Sage Handbook on Comparative Studies in Education: Practices and experiences in student schooling and learning*. SAGE.

Purves, A.C. (1987). The evolution of the IEA: A memoir. *Comparative Education Review, 31*, 10-28.

Robitzsch, A. (2020). L_p loss functions in invariance alignment and haberman linking with few or many groups. *Stats, 3,* 246-283. https://doi.org/10.3390/stats3030019

Rosenshine, B. (1976). Classroom instruction. In N.L. Gage (Ed.), *The psychology of teaching methods*. University of Chicago Press.

Rowan, B. (2002). The ecology of school improvement: Notes on the school improvement industry in the United States. *Journal of Educational Change*, *3,* 283-314.

https://doi.org/10.1023/A:1021277712833

Rutkowski, L., & Rutkowski, D. (2019). Methodological challenges to measuring heterogeneous populations internationally. In L.E. Suter, E. Smith, & B.D. Denman (Eds.), *The SAGE handbook of comparative studies in education*. SAGE.

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. https://doi.org/10.1177/0013164413498257

Rutkowski, L., & Svetina, D. (2017). Measurement invariance in international surveys: Categorical indicators and fit measure performance. *Applied Measurement in Education, 30,* 39-51. https://doi.org/10.1080/08957347.2016.1243540

Sachisthal, M.S.M., Jansen, B.R.J., Peetsma, T.T.D., Dalege, J., van der Maas, H.L.J., & Raijmakers, M.E.J. (2019). Introducing a science interest network model to reveal country differences. *Journal of Educational Psychology, 111*, 1063–1080. https://doi.org/10.1037/edu0000327

Santagata, R. (2005). Practices and beliefs in mistake-handling activities: A video study of Italian and US mathematics lessons. *Teaching and Teacher Education, 21,*491-508. https://doi.org/10.1016/j.tate.2005.03.004

Scheerens, J. (2017). *Opportunity to learn, curriculum alignment and test preparation: A research review.* Springer. https://doi.org/10.1007/978-3-319-43110-9

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.* https://doi.org/10.3389/fpsyg.2016.00110

Schwarz, N. (2007). Cognitive aspects of survey methodology. *Applied Cognitive Psychology, 21,* 277–287. https://doi.org/10.1002/acp.1340

Schwarz, N., Oyserman, D., & Peytcheva, E. (2010). Cognition, communication, and culture: Implications for the survey response process. In J.A. Harkness, M. Braun, B. Edwards, T.P. Johnson, L. Lyberg, P. Ph. Mohler, B.-E. Penell, & T.W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts.* Wiley. https://doi.org/10.1002/9780470609927.ch10

Scott, J. (1988). Social network analysis. *Sociology*, *22*, 109-127. https://doi.org/10.1177/0038038588022001007

Steenkamp, J-B.E.M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25,* 78-90. https://doi.org/10.1086/209528

Stigler, J.W., & Hiebert, J. (1999). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom.* The Free Press.

Stringfield, S.C., & Slavin, R.E. (1992). A hierarchical longitudinal model for elementary school effects. In B.P.M. Creemers & G.J. Reezigt (Eds.), *Evaluation of educational effectiveness*. ICO.

Tabulawa, R. (2003). International aid agencies, learner-centred pedagogy and political democratisation: A critique. *Comparative Education*, 39, 7-26. https://doi.org/10.1080/03050060302559

Täht, K., & Must, O. (2013). Comparability of educational achievement and learning attitudes across nations. *Educational Research and Evaluation, 19,* 19-38*.* http://dx.doi.org/10.1080/13803611.2012.750443

Tobias, S. & Duffy, T.M. (Eds.) (2009). *Constructivist instruction: Success or failure?* Routledge.

Tobin, J., Hsueh,Y., & Karasawa, M. (2009). *Preschool in three cultures revisited.* University of Chicago Press.

Tourangeau, R. (1984). Cognitive science and survey methods: A cognitive perspective. In T. Jabine,

M. Straf, J. Tanur, & R. Tourangeau (Eds.), *Cognitive aspects of survey methodology: Building a bridge between disciplines*. National Academy Press.

Tourangeau, R., Rips, L.J., & Rasinski, K. (2000). *The psychology of survey response.* Cambridge University Press. https://doi.org/10.1017/CBO9780511819322

Uskul, A.K., & Oyserman, D. (2006). Question comprehension and response: Implications of individualism and collectivism. In B. Mannix, M. Neale, & Y. Chen (Eds.), *Research on managing groups and teams: National culture & groups.* Elsevier Science Press.

van Borkulo, C.D., Boschloo, L., Borsboom, D., Penninx, B.W.J.H., Waldorp, L.J., & Schoevers, R.A. (2015). Association of symptom network structure with the course of depression. *JAMA Psychiatry*, *72*, 1219–1226. https://doi.org/10.1001/jamapsychiatry.2015.2079

van Borkulo, C.D., Epskamp, S., Jones, P., Haslbeck, J., & Millner, A. (2016). *Network comparison test: Statistical comparison of two networks based on three invariance measures.* Retrieved from https://cran.r-project.org/web/packages/NetworkComparisonTest/index.html

van Borkulo, C.D., van Bork, R., Boschloo, L., Kossakowski, J.J., Tio, P., Schoevers R.A., Borsboom, D., & Waldrop, L.J. (2017). *Comparing network structures on three aspects: A permutation test.* https://doi.org/10.13140/RG.2.2.29455.38569

van de Gaer, E., Grisay, A., Schulz, W., & Gebhardt, E. (2012). The reference group effect: An explanation of the paradoxical relationship between academic achievement and self-confidence across countries. *Journal of Cross-Cultural Psychology, 43,* 1205–1228. https://doi.org/10.1177/0022022111428083

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

Vandenberg, R.J., & Lance, C.E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3,* 4–70. https://doi.org/10.1177/109442810031002

van de Vijver, F.J.R. (1998). Towards a theory of bias and equivalence. In J.A. Harkness (Ed.), *Cross-cultural survey equivalence.* ZUMA

van de Vijver, F.J.R, Avvisati, F., Davidov, E., Eid, M., Fox, J.-P., Le Donne, N., Lek, K., Meuleman, B., Paccagnella, M., & van de Schoot, R. (2019). Invariance analyses in large-scale studies. *OECD Education Working Papers*. OECD Publishing. https://doi.org/10.1787/254738dd-en

van de Vijver, F.J.R., & He, J. (2016). Bias assessment and prevention in noncognitive outcome measures in context assessments. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Assessing contexts of learning. Methodology of educational measurement and assessment.* Springer. https://doi.org/10.1007/978-3-319-45357-6_9

van de Vijver, F.J.R., & Leung, K. (1997). Methods and data analysis of comparative research. In J.W. Berry, Y.H. Poortinga, & J. Pandey (Eds.), *Handbook of cross-cultural psychology*. Allyn & Bacon.

van de Vijver, F.J.R., & Matsumoto, D. (2011). Introduction to the methodological issues associated with cross-cultural research. In D. Matsumoto & F.J.R. van de Vijver (Eds.), *Cross-cultural research methods in psychology*. Cambridge University Press.

Varnum, M.E.W., Grossmann, I., Kitayama, S., & Nisbett, R.E. (2010). The Origin of cultural differences in cognition: The social orientation hypothesis. *Current Directions in Psychological Science, 19,* 9–13. https://doi.org/10.1177/0963721409359301

Vygotsky, L.S. (1978). *Mind in society: Development of higher psychological processes.* Harvard University Press.

Welkenhuysen-Gybels, J.G.J., van de Vijver, F.J.R., & Cambré, B. (2007). A comparison of methods for

the evaluation of construct equivalence in a multi-group setting. In G. Loosveldt, M. Swyngedouw, & B. Cambré (Eds.), *Measuring meaningful data in social research,* 357-72. Acco.

Willis, G.B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. SAGE.

Willis, G.B. (2015). The practice of cross-cultural cognitive interviewing. *Public Opinion Quarterly, 79,* 359–395. https://doi.org/10.1093/poq/nfu092

Willis, G.B., & Miller, K. (2011). Cross-cultural cognitive interviewing: Seeking comparability and enhancing understanding. *Field Methods, 23,* 331-341. https://doi.org/10.1177/1525822X11416092

Appendix

Appendix A: Manuscript 1

Material from:

Fischer, J., Praetorius, A.-K., & Klieme, E. The impact of linguistic similarity on cross-cultural comparability of students' perceptions of teaching quality. *Educational Assessment, Evaluation and Accountability (2019)31:201–220.* https://doi.org/10.1007/s11092-019-09295-7

Educational Assessment, Evaluation and Accountability (2019) 31:201–220
https://doi.org/10.1007/s11092-019-09295-7

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]

Received: 25 October 2018 / Accepted: 2 April 2019 / Published online: 15 April 2019

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.

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).

[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.

2.3 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 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."

[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).

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.

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 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

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

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

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

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).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

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.

Appendix B: Manuscript 2

Fischer, J., Klieme, E., Praetorius, A.-K., & Jinjie, X. (submitted). Understanding lack of equivalence in cross-cultural measurements of teaching quality: Students' interpretations of student support items in Germany and China. *Submitted to Teaching and Teacher Education*.

Understanding lack of equivalence in cross-cultural measurements of teaching quality: students' interpretations of student support items in Germany and China

Jessica Fischer[1], Eckhard Klieme[1], Anna-Katharina Praetorius[2], Xu Jinjie[3]

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

[2] University of Zurich, Switzerland

[3] Research Institute for International and Comparative Education, Shanghai Normal University, China

Declarations of interest: The authors declare that they have no conflict of interest.

Author Note

Correspondence concerning this article should be addressed to Jessica Fischer, DIPF | Leibniz Institute for Research and Information in Education, Rostocker Straße 6, 60323 Frankfurt am Main, Germany. E-mail: jessica.fischer@dipf.de

Abstract

There is a growing interest in studying teaching quality, a powerful predictor of learning outcomes, in cross-cultural surveys. Yet, teaching quality measures often work differently across cultures and thus scores cannot be compared validly. Limited effort has been made to understand the lack of comparability, which we seek to shed light on. We conducted cognitive interviews with students in Germany and China and comparatively analysed their interpretations of PISA student support items. We found culture-specific interpretative variations for statistically non-comparable and comparable items, which were linked to translation differences, item characteristics, and culture-specific teaching traditions and definitions of student support.

Keywords

student support; teaching quality; cross-cultural comparability; survey response process; cognitive interviewing; qualitative content analysis

1 The challenge of measuring teaching quality cross-culturally

Teaching quality is one of the most powerful predictors of student learning outcomes (Hattie, 2009) and it thus has been investigated extensively over the last decades. To conceptualise teaching quality, these investigations used varying frameworks. One prominent framework is the framework of the Three Basic Dimensions, comprising the dimensions classroom management, student support, and cognitive activation (Klieme, Pauli, & Reusser, 2009). *Classroom management* refers to maximising learning time and covers sub-dimensions such as the effective handling of disruptions and use of time, classroom discipline, clarity of rules, and monitoring. *Student support* encompasses teaching aspects with a focus on students, such as supporting social relationships, autonomy, and competence. *Cognitive activation* summarizes aspects that promote students' higher level thinking, such as the use of challenging tasks, exploring and activating prior knowledge, supporting metacognition, or eliciting student thinking (Praetorius, Klieme, Herbert, & Pinger, 2018).

The framework was originally developed based on mathematics instruction in German-speaking countries, and a comparable factor structure and measurement was empirically supported in Germany across subjects, school types, and grade levels (Fauth, Decristan, Rieser, Klieme, & Büttner, 2014; Wisniewski, Zierer, Dresel, & Daumiller, 2020). Meanwhile, the teaching quality dimensions have gained considerable international attention, and empirical studies have been published internationally (e.g., Dorfner, Förtsch, & Neuhaus, 2018; Fauth et al., 2014; 2019; Nilsen & Gustafsson, 2016; Praetorius, Pauli, Reusser, Rakoczy, & Klieme, 2014; Yi & Lee, 2017). The teaching quality dimensions have been used in multiple international large-scale assessments (e.g., TALIS: Vieluf, Kaplan, Klieme, & Bayer, 2012; PISA: Scherer, Nilsen, & Jansen, 2016; TIMSS: Nilsen & Gustafsson, 2016). However, even though the dimensions are theoretically conceptualized as being generalizable across education systems (Praetorius et al., 2018b), the valid

cross-cultural measurement of teaching quality is a huge challenge. In a study using PISA 2012 data, a limited cross-cultural comparability of teaching quality measures across 15 education systems was identified (Fischer, Praetorius, & Klieme, 2019). The level of comparability varied between dimensions and items and was particularly low across culturally and linguistically diverse groups. Likewise, other studies have statistically demonstrated a limited comparability of teaching quality measures, so that means of teaching quality cannot be compared validly across education systems (Desa, 2014; He, Buchholz, & Klieme, 2017; He & Kubacka, 2015; Scherer et al., 2016; Schulz, 2005; van de Grift, 2014).

Despite its critical relevance, limited effort has been made to understand why questionnaire measures of teaching quality often work differently across education systems. Cultural psychology has consistently demonstrated that cultures vary in how information is processed (Varnum, Grossmann, Kitayama, & Nisbett, 2010), which may systematically shift the meaning of items as a function of the context in which the question is being asked. Thus, understanding similarities and differences in the cognitive processing of teaching quality items can provide valuable insight into potential sources of incomparability, and consequently, can aid comparable measurement. To fill the current research gap, the present study evaluates the impact of culture on the interpretation of student support items, as no other study has investigated the cognitive processes for student support items so far. Given the three dimensions, a culture-specific interpretation seems particularly likely for items measuring student support (see Section 1.4). Therefore, as persuasive differences in basic cognitive processes have been found for Western versus East-Asian respondents (Schwarz, Oyserman, & Peytcheva, 2010), we conducted cognitive interviews with students from Germany and China (Shanghai), and comparatively analysed the interpretative patterns for statistically comparable and non-comparable student support items.

In the following, we first summarize the cognitive stages respondents engage in while answering survey items (Section 1.1) and highlight the potential impact of culture (Section 1.2). Afterward, we describe cross-cultural cognitive interviewing, a method to detect cultural variations (Section 1.3), leading to our research questions and hypotheses focusing on student support (Section 1.4).

1.1 Stages involved in answering survey questions and potential problems

The interdisciplinary research field "Cognitive Aspects of Survey Methodology" (CASM) studies the cognitive and communicative processes of respondents while answering survey items with a focus on the *cognitive validity*, i.e., the degree that cognitive processes mirror those intended by the researcher (Karabenick et al., 2007). In the field, there is shared agreement that answering items involves a series of cognitive stages. One prominent conceptualisation is the Model of Response Process which proposes four stages and corresponding mental processes (Tourangeau, 1984; Tourangeau, Rips, & Rasinski, 2000). *Comprehension* is the first stage, the interpretation of the item. It encompasses mental processes such as attending to the item and accompanying instructions as well as inferring the meaning of the item. This helps respondents to understand which information they ought to provide. Problems may arise if respondents misinterpret the item, for instance by missing part of the item or because of complicated or unfamiliar wordings. *Information retrieval* involves recalling relevant information from long-term memory or based on whatever information is accessible at the moment of survey response. Among others, adopting a retrieval strategy or recollecting memories are relevant mental processes. Accuracy and completeness are affected by several aspects, such as the number and quality of cues in the item or the source of memory (e.g., own or second-hand knowledge). The *judgment* stage comprises processes that respondents use to combine the retrieved information with the item in order to form

a "private" judgement. Errors can occur if respondents draw erroneous inferences (e.g., judgments are based on misinterpretations of the item). The last stage is the actual *reporting and response selection* – mapping the judgment to fit the answering scale and editing the response. Respondents may have difficulties with the set of answering categories or differ in their approaches of selecting a category. Further, they might not be willing to answer the question truthfully due to social desirability, hence a mismatch may occur between the respondent's judgment and a selected answering category. Item interpretation, in particular, plays a central role in the response process. The stages are conceptualized as sequential, thus, the respondents' interpretation affects the content of retrieved information, which may shape the respondent's judgment and finally the actual response. Further, respondents may have an answer ready after interpreting the item, and thus skip information retrieval and judgment formation (Schwarz, 2007; Tourangeau et al., 2000).

With regard to teaching quality measures, research is still scarce as to whether the students' cognitive processes mirror those intended by the researchers. First empirical evidence by Lenske (2016) and Lenske and Praetorius (2020) suggests that German students often do not process teaching quality items in a valid manner. This was foremost linked to errors occurring in the *comprehension stage*: In both the cited studies, hardly any of the interviewed students on the primary and secondary school level interpreted the evaluated items measuring classroom management and cognitive activation as intended. For cognitive activation, the items often did not even measure the targeted construct, and thus it was impossible to draw valid conclusions based on student ratings. As most frequent sources of errors, Lenske (2016) identified misinterpretations of key terms, complex items, or unfamiliar wordings. However, many students associated the item content with relevant experiences in the classroom (yet some students invented some aspects, *information retrieval stage*), and were able to select an appropriate answering category (*response stage,* Lenske, 2016). Yet, none of the referenced studies adopted a cross-cultural comparative

perspective and to our knowledge, no study has investigated the cognitive validity of student responses for student support items so far.

1.2 The impact of culture on survey responses

Culture-specific variations in the response process can be a serious threat to the comparable measurement of teaching quality. To date, research has paid limited attention to cultural differences in survey responding and to our knowledge, no study has investigated the impact of culture on the responses process for teaching quality measures. However, there is first evidence that culture may influence the response process at all stages elaborated above (Schwarz, 2007; Schwarz et al., 2010; Uskul & Oyserman, 2006).

Comprehension stage. In cross-cultural measurement of teaching, the involvement of various languages probably poses the most serious threat to a comparable item interpretation. One source of interpretative variations are translation errors. As a consequence, translations are not functionally equivalent, which can alter the meaning of an item. Fitzgerald and colleagues (2011) identified several types of translation errors that can compromise the equivalence between linguistic versions, such as omitted or unfamiliar words. Further, translations can differ with regard to complexity. Simpler formulations are expected to enhance the cognitive validity of survey items. If, however, the term in a certain language is more complex than in others, the likelihood of misjudgements increases for that group (van de Vijver & Leung, 1997). Culture-specific designs of instruction can further endanger a comparable item interpretation. Based on a lexicon study, Mesiti and Clarke (2017) point out that teaching practices may be specific to individual education systems. Consequently, there are no words to describe them in other languages, which can lead to inaccurate translations. For cross-cultural measures of teaching and learning, empirical studies have identified differential effects of language and translation on item difficulty (for an overview

see Hopfenbeck et al., 2017). According to Grisay and colleagues (2007), translations are more accurate for Western compared to Middle Eastern or Asian countries, as those languages are more similar to the source languages. Still, even linguistic equivalence does not safeguard against differential item interpretations. Language is a powerful indicator of cultural identity and can predispose certain choices of interpretation. Pedagogical histories and norms of practices are enshrined in the language that is used to describe classroom phenomena. Mesiti and Clarke (2017) demonstrated that words describing experiences in classroom can have a different meaning across education systems, which can be a threat to the construct validity of measures. For instance, "participation" in China means choral response, but student talk in Australia. Thus, the same word may have a different connotation depending on the cultural context. Similarly, Benítez and Padilla (2014) identified variations in the interpretation of key words between Spanish and U.S. respondents for items on the value of science, which the authors partly linked to differences in the social and health systems. Besides linguistic issues, cultural differences with regard to the sensitivity to instrument characteristics (i.e., item order effects) can result in interpretative variations. For instance, compared to Western respondents, East-Asian respondents are less likely to give redundant information if they have already provided similar answers to preceding items/questions (Schwarz et al., 2010).

Information retrieval and judgment formation stage. Once respondents have determined the information they are supposed to provide, they need to recall relevant memory and form a judgment. Cultural groups vary in content and organization of memory, as well as regarding what they attempt to retrieve, and how they organize information in a narrative form. Three main differences have been identified: the focus on self versus others, depth versus breadth of memory, and internal state versus context. For instance, when reporting about experiences, Western respondents more frequently refer to the self, focus on isolated aspects of personal interest, and make more references

to their internal states, compared to Chinese respondents. Chinese respondents tend to focus on others, to report about events as a whole, and hardly refer to personal desires or emotions (Schwarz et al., 2010). For teaching quality items, students have to recall relevant experiences and behaviours in classroom to form a judgment. Hence, depending on the cultural context, students may differ in their perceptions of those experiences, remember and focus on different aspects, and vary in the extent that their perceptions are influenced by internal states. Moreover, the associated experiences and behaviours themselves can be expected to differ. Researchers have repeatedly demonstrated that teaching can be described as a cultural activity that is only generalizable across education systems to a limited extent. Accordingly, education systems show distinctive instructional features and priorities of classroom practice, thus experiences that students make in classrooms across the world are not necessarily the same (Clarke, Keitel, & Shimizu, 2006; Stigler & Hiebert, 1999). By understanding which experiences and behaviours are associated with the item content, we can gain evidence of whether measures show an adequate construct representation across education systems. However, research is still scarce with regard to items measuring teaching quality in a cross-cultural context.

Editing and response selection stage. In addition to the item content, the respondents' cultural background can shape the actual response. Social desirability, meaning answering in culturally-sanctioned ways (Crowne & Marlowe, 1964), can be seen as universal as everyone strives to represent themselves in a positive light. Yet, the acceptable strategies and the content that is considered favourable vary across cultures, depending on the preferred evaluation and presentation of the self (Schwarz et al., 2010). These variations have been linked to culture-specific response styles, thus, a systematic tendency to respond to items on some basis other than the item content (Paulhus, 1991). Among others, culture-specific response tendencies may result in a different use of rating scales. So far, the contrast between Western versus East-Asian respondents has received

the most attention. Compared to Western respondents, impression management is more important for East-Asian respondents, leading to a preference for midpoint responding (i.e., choosing middle response categories) and acquiescence (i.e., tendency to agree). Western respondents are less likely to engage in socially desirable responding, yet they show a tendency to self-enhancement, i.e., extreme responding (e.g., Harzing, 2006). Statistical methods to detect and control for culturally-shaped response tendencies (e.g., van Herk, Poortinga, & Verhallen, 2004) and instruments to correct for responses that might be affected by incomparability have been developed. For instance, He and colleagues (2017) demonstrated that, in comparison to Likert-Type scales, anchoring vignettes measuring classroom management and student support were less likely to be biased by culture-specific response tendencies.

To summarize, culture can impact the survey response process and shift the meaning of items and answers. Consequently, different scores might represent true differences in teaching quality, differences in the response process, or an unknown combination of both. Therefore, it is crucial to unfold the role played by culture in the response process.

1.3 Detecting cultural variations: cross-cultural cognitive interviewing

Cognitive interviewing (oftentimes referred to as cognitive labs or cognitive pretesting) is one of the most prominent methods to evaluate the cognitive validity of survey items (Koskey, Karabenick, Woolley, Bonney, & Dever, 2010). If cognitive interviewing is conducted in a cross-cultural context, it is referred to as cross-cultural cognitive interviewing (CCCI) (Willis & Miller, 2011). CCCI can aid the improvement of measurement across groups and the substantive interpretation of data by providing insight into how survey items are approached, consumed, and digested (Behr, Braun, Kaczmirek, & Bandilla, 2014). Ideally, CCCI provides evidence that there

are no differences in the survey response processes across groups. Otherwise, it has the potential to identify sources of variation, occurring in different cognitive stages.

Two major techniques for prompting respondents to verbalise their response processes can be distinguished. *Think-aloud* (also concurrent verbalisation) requires the respondents to verbalize their thoughts during each item response, whilst the interviewer interjects as little as possible. Think-aloud is expected to hardly be affected by interviewer bias and to represent literal reflections of the actual thought process, yet it heavily depends on the respondent's ability to perform the think-aloud. *Verbal probing* is characterized by an interviewer asking follow-up questions to elicit additional information during (concurrent probing) or after the respondent has completed the questionnaire (retrospective probing) (Willis, 2005). Probes help to focus on issues important to the researcher and can be divided into general (e.g., "*Did you understand the question?")* and specific probes. Specific probes are more frequently used and are designed to capture a specific cognitive stage (for examples of different kinds of probes see e.g., Lietz, 2017).

CCCI may be a compelling means to evaluate the comparability of survey items, yet it faces several challenges due to the intercultural context. Not only do the members of the research team differ with regard to their cultural background and native language (which can lead to communication problems within the team), interviewing in different cultural contexts and translating instruments and audio data can bias results (Willis & Miller, 2011). Thus, to make CCCIs truly comparable, Miller and colleagues (2011) suggest standardizing as many aspects as possible without limiting the advantage of CCCI's interpretive qualities. The authors recommend semi-structured verbal probing: in addition to a set of pre-scripted standardized probes, interviewers have some flexibility to pose additional questions matched to the situation and cultural context, while exclusively relying on standardized probes may fail to capture the response process

across groups adequately. It is equally important to focus on strategies that improve the comparability of CCCI itself, for instance discussing and analysing results jointly with all collaborators of a study in order to avoid a culturally-biased interpretation of the data.

1.4 The present study

The aim of this study is to investigate culture-specific similarities and differences in the cognitive processing of teaching quality items and to identify potential sources of the limited comparability of teaching quality measures.

We focus on the comprehension and information retrieval stage as they are important stages in the response process that are likely shaped by culture. Further, compared to the other cognitive stages of survey responding, culture-specific response tendencies *(response selection stage)* are comparatively well researched (see Sections 1.1. and 1.2).

Conducting CCCIs is very time consuming so we focus on one dimension of teaching quality, namely student support. This decision is based on several considerations: 1) No other study has investigated the cognitive validity of student responses for student support items so far. 2) A culturally-shaped understanding of what is seen as supporting behaviour and the associated experiences in classroom seem likely, depending on the education systems' social orientation (e.g., power distance, thus, the degree that less powerful members of institutions except the unequal power distribution, Hofstede, 2001; see Vieluf, Kunter, & van de Vijver, 2013), preferred instructional approaches (e.g., a more teacher or student-directed approach, see Fischer, He, & Klieme, 2020), and instructional goals (e.g., enhancing motivation or values versus knowledge and skills, see Praetorius et al., 2018a). 3) Moreover, when reported by students, perceptions of student support are shaped by individual cognitions and emotions, which are known to impact survey response (Karabenick et al., 2007). As persuasive cognitive differences have been found for East-

Asian versus Western respondents (see Section 1.2 and Schwarz et al., 2010), we are interested in: *how students from Germany and China (Shanghai) interpret student support items and which experiences in classroom are associated with the item content.*

Linking qualitative differences in item interpretation to the statistical analyses of measurement equivalence, we further hypothesize: For *statistically comparable* items, item interpretation and associations hardly differ between students from Germany and China (Shanghai) (*Hypothesis 1*) while for *statistically non-comparable* items, interpretation differs more strongly (*Hypothesis 2*).

2 Methods

2.1 Selecting items for cross-cultural cognitive interviews

Measures. In this study, we evaluate student support items that are frequently used in the field of teaching quality (see Praetorius et al., 2018b), in a cross-cultural context for instance in the Program for International Student Assessment (PISA). Since 2000, PISA has measured 15-year old students' perceptions of student support with five items on a 4-point Likert-type scale ranging from "every lesson" to "never or hardly ever" (OECD, 2014, for the English source, for the German and Chinese (Shanghai) versions see Table 1). In this study, we use items framed for mathematics instruction, as implemented in PISA 2012. Values of Cronbach's alpha of .85 (Germany) and .84 (China, Shanghai) indicate good scale reliability. To ensure a reasonable response burden, interview duration, and amount of narrative data for analysis, we aimed at selecting three of the originally five PISA items to be evaluated as part of the CCCIs. As we expect item interpretation to exhibit a low variation for statistically comparable items (*Hypothesis 1*), but rather high variations for statistically non-comparable items (*Hypothesis 2*), we aimed at selecting one comparable and two non-comparable items based on PISA 2012 data for Germany and China (Shanghai).

Table 1

PISA student support items – English source, German, and Chinese (Shanghai) version

Source Version	German Version	Chinese (Shanghai) Version
How often do these things happen in your mathematics lessons?		
TS01: The teacher shows an interest in every student's learning.*	TS01: Unsere Lehrerin/unser Lehrer interessiert sich für den Lernfortschritt jeder einzelnen Schülerin/jedes einzelnen Schülers.	TS01: 老师关注每个学生的学习状态。
TS02: The teacher gives extra help when students need it.		
TS04: The teacher helps students with their learning.*	TS04: Unsere Lehrerin/unser Lehrer unterstützt uns beim Lernen.	TS04: 老师帮助学生学习。
TS05: The teacher continues teaching until the students understand.*	TS05: Unsere Lehrerin/unser Lehrer erklärt etwas so lange, bis wir es verstehen.	TS05: 老师会一直讲解，直到学生理解为止。
TS06: The teacher gives students an opportunity to express opinions.		

Note. *Items selected for cognitive interviews. Response categories: 1= Every lesson, 2= Most lessons, 3= Some lessons, 4= Never or hardly ever

Analysis strategy. For detecting statistical comparability, we applied the Alignment-method by Asparouhov and Muthén (2014) to the PISA 2012 data on student support in Germany and China (Shanghai). Alignment does not require full data comparability, which is hardly satisfied in a cross-cultural context, but estimates a solution that minimizes the differences between the parameters (factor loadings and intercepts) across groups with a procedure that is similar to rotation in exploratory factor analysis. The underlying assumption is that most parameters are equivalent, and a minority are not. Besides checking equivalence, Alignment flags statistically comparable and non-comparable items (with regard to factor loadings and intercepts), which renders the method suitable to select both comparable and non-comparable items for the CCCIs. Students with missing data on all items were excluded from analysis. To avoid different model contributions due to varying sample sizes, a subsample of 1.000 students was drawn according to final student weights.

We accounted for the hierarchical structure of PISA data (i.e., students nested in schools and education systems) (OECD, 2014). Analyses were conducted with Mplus Version 8.4 (Muthén & Muthén, 1998-2017).

2.2 Evaluating item interpretation: Cross-cultural cognitive interviews

To evaluate the selected items, we conducted CCCIs in Germany and China (Shanghai). Sampling, instruments, and analysis method are described in the following.

Sampling. The participants were chosen to mirror the characteristics of participants in the PISA 2012 study. Three secondary schools were selected in Germany (in two states) and China (Shanghai). Based on differences in education system, one low, one medium, and one high performing secondary school was selected in China (Shanghai), whereas school performance was unknown in Germany. Within schools, seven mathematics teachers per education system who agreed to participate[1] selected two students of their mathematics class (14 students per system, 28 in total) based on the following criteria: 1) students had to be about 15 years old (equal to the PISA sample), 2) the native language had to be German in Germany and Mandarin in China (Shanghai) (as we targeted the interpretation of mainstream students), 3) roughly half of the students should be female/male, 4) high and low performing students had to be similarly represented, and 5) students had to have good verbalizing skills. In Germany, 57 percent of the students were male (PISA 2012: 51% male), had an average last math grade of 3 (equals "satisfying"), and 175 books at home on average (PISA 2012: 170 books). In China (Shanghai), 57 percent of the sampled students were male (PISA 2012: 49% male), had an average last math grade of 123 (equals "good"), and an average of 80 books at home (PISA 2012: 100 books).

[1] For some schools more than one teacher agreed to participate.

Instruments and methods of data collection. After completing a context-questionnaire assessing gender, date of birth, number of books at home, and last mathematics grade, students verbally completed a questionnaire ("think-aloud") consisting of the three selected items, worded identically to the German/Shanghai PISA 2012 student questionnaire. Subsequently, they answered follow-up questions during a semi-structured retrospective probing. However, most students had difficulties to freely think aloud and thus the interpretative value of the think aloud was rather low. Hence, we only describe the probing in the following.

Table 2

Probing-protocol for the cognitive interviews

Stages of the model of response process	Obligatory probes	Optional probes
	Introduction: 1. For [*item y*], you chose [*answering category xx*].	
Item comprehension incl. key terms (Stage 1)	Paraphrasing: 2. Can you repeat [*item y]* in your own words? 3. What does [*key term item y*] mean to you?	Elaboration/ Expansion: a) Can you explain that in more detail? b) What does that mean?
Associations with experiences during instruction considered for item interpretation (Stage 2)	Elaboration: 4. Can you give me an example, what a teacher does in math lessons, if [*item y]*.	4a) Expansion: Can you give me more examples? 4b) How do <u>you</u> notice, that *[item y]*. Elaboration/ Expansion: a) Can you explain that in more detail? b) What does that mean?

Note. The protocol was applied for every item that was evaluated as part of the CCCIs. *[item y]*=insert wording of individual item; *[answering category xx]*=insert chosen answering category; [*key term item y*]= insert key term of item

The semi-structured probing protocol (see Table 2) consisted of a set of pre-scripted standardized probes targeting in-depth information on 1) item interpretation including key terms and 2) associations of the item content with experiences in classroom. The interviewers had the flexibility to vary the sequence of the standardized probes and to ask additional spontaneous probes if they thought them to be necessary in order to understand specific statements, either because they reflected culture-specific situations or they needed further clarification. The probes were developed in German, translated into Chinese by a professional translation company, and adapted by the Chinese research team. Prior to the actual data collection, the probing protocol was pretested. To eliminate interviewer effects and to standardize the interviews most optimally, the same (German) interviewer probed the German (together with two assistants) and Chinese (Shanghai) students. In China (Shanghai), two members of the Chinese research team (in addition to the German interviewer and interpreter) were present during the probing in order to adjust the probes to the cultural context (if necessary). Yet, due to time restrictions set by the participating schools, 6 of the 14 interviews had to be conducted by the Chinese research team. With no time limit specified, the average duration was 12 minutes in Germany (range=05:47-16:01minutes) and 7 minutes in China (Shanghai) *without* interpreter (range=04:45-09:15 minutes) and 13 minutes *with* interpreter (range=09:45-16:26 minutes). The on average shorter duration of the Chinese interviews also reflects the tendency of Chinese respondents to give short, non-redundant answers (Uskul & Oyserman, 2006). The interviews were audio-recorded and transcribed in German (according to rules by Kuckartz, 2018), the Chinese data by a native Chinese translator with expertise in math instruction.

Development of the coding system. To analyse the narrative data, we applied computer-aided Qualitative Content Analysis (Kuckartz, 2018). We developed our coding system in several cycles. First, two (German) coders independently developed a coding system based on the German data

(data-driven), meaning the (main- and sub-)codes were generated and revised as long as new information was found within the data during the content analysis process. Afterward, the two coders compared and discussed their coding systems and a joint system was developed with an expert's input on teaching and learning (adding deductive codes based on existing theories and literature), which was then used to re-code the German data. In a second step, one (German) coder applied the coding system, which was developed based on the German data, to the Chinese data and codes were added or existing codes revised if necessary. The resulting coding system was then applied by a second (German) coder to code the Chinese (Shanghai) data and again revised based on feedback provided by the second coder, the Chinese research team, and experts on qualitative research and instruction. The final coding system was then used to re-code all interviews.

Analysis procedure. Miller and colleagues (2011) recommend to analyse CCCI data following three subsequent steps, which we adhered to (we only report Step 3): analysing 1) *within an interview* in order to a) understand how respondents interpret items and to b) identify differences and similarities across items of a scale; 2) *across interviews* of respondents belonging to the *same group* in order to identify common or different interpretive patterns across items and interviews; 3) and *across groups* in order to identify differences and similarities across items and groups. Coding and analysis was conducted with MAXQDA Version 2018.2 (VERBI Software, 2018).

3 Results

3.1 Selecting items for cross-cultural cognitive interviews

For all five student support items, the factor loadings were equivalent (see Table 3). Based on intercept (non-)equivalence, we selected Item TS01 ("The teacher shows an interest in every student's learning") and TS04 ("The teacher helps students with their learning") as the two non-comparable items for the CCCIs. Thus, according to the statistical analysis, Chinese and German

students differ in their understanding and response to items TS01 and TS04. From the comparable items (equivalent intercepts for Germany and China, Shanghai), we selected TS05 ("The teacher continues teaching until students understand"). As more than one item was statistically comparable, the selection was based on theoretical considerations, such as the goal to cover wide aspects of the construct.

Table 3

Comparability of PISA student support items for Germany and China (Shanghai) and sources of differences in each item identified by cognitive interviews

Item	Factor loadings	Item Intercepts	Identified sources of differences		
			Translation	Item complexity and content	Cultural context
TS01	GER QCN	**(GER) (QCN)***	X	X	X
TS02	GER QCN	GER QCN			
TS04	GER QCN	**(GER) (QCN)***			X
TS05	GER QCN	GER QCN*		X	X
TS06	GER QCN	GER QCN			

Note. Measurement invariance testing with alignment. Parentheses indicate non-equivalent factor loadings/ intercepts for that specific group. GER=Germany; QCN=China (Shanghai). *Items selected for cognitive interviews. "X" indicates the presence of that source of differences at the qualitative level for each item

3.2 Evaluating item interpretation: Cross-cultural cognitive interviews

In the following, we describe the central findings of the probing separately for each item. The main codes (in "bold print" in the text) and sub codes that were used to identify interpretative patterns and their frequencies are presented in Table 4 and identified sources of difference in Table 3. In total, the number of assigned codes was higher for the German data (284 codings) compared to the Chinese (Shanghai) data (141 codings) and the variety of assigned codes was higher for Germany, indicating that the Chinese (Shanghai) students showed less variation with regard to item

interpretation, but also provided shorter answers on average (which is in line with the on average shorter interview duration).

Table 4

Final coding system for the cognitive interviews and frequencies per item

	Item TS01: interest in student's learning		TS04: help with learning		TS05: explaining until students understand	
Main Sub-codes	GER	QCN	GER	QCN	GER	QCN
Teachers' goal						
Understanding	6	2	1	-	-	-
Learning success	3	-	-	-	-	-
Attention	-	10**	-	-	-	-
Reference group						
Majority	3	-	1	-	5**	-
High-achieving students	-	-	-	-	-	4**
Content-focused teaching						
Additional learning opportunities	5**	-	11**	-	-	-
Application math methods	-	-	1	10**	-	-
Monitoring						
Understanding	10	5	9	5	6**	-
Attention	-	9**	-	-	-	-
Teaching practices						
Remedial activities	12**	4	13	13	14	14
Fostering attention	1	5	1	-		-
Adaptivity	2	1	-	3	3	-
Learning with other students	3	-	4**	-	7**	-
Time management						
Flexible	2	-	1	-	6**	-
Fixed	-	-	-	-	2	10**
Autonomous student learning	-	-	4**	-	-	-
Socio-emotional experience	1	-	4**	-	4**	-

Note. Main-codes are in "bold print". GER=Germany; QCN=China (Shanghai). Sub-codes are counted only once per interview. ** Difference is significant based on Chi-Square tests

3.2.1 Item TS01: The teacher shows an interest in every student's learning

Statistically, Item TS01 was not comparable. Of all items we evaluated as part of the CCCIs, TS01 showed the most distinct interpretative variations. As we found during our empirical work, this was based on differences in translation of the key term "learning". The German term "Lernfortschritt" means learning *progress,* which the students referred to as processes that can be achieved in a short or extended time, such as personal growth, increase in knowledge (individual and class level), eliminating misunderstandings, or instructional progress. The Chinese term "学习状态" means learning *state*, which the students referred to as motivational, emotional, and cognitive learning dispositions. It was described as an image of a students' learning condition at a definite point in time, which can be less stable and change direction (e.g., positive or negative mood),

> German student, 05DSMI26: *Learning progress means that you know more after the lesson (...) you understand the theory and you also know how to use it to solve the corresponding mathematical tasks*.
>
> German student, 05DSNI22: *Learning progress means, how far you have progressed with a specific topic in school. If, for instance, you are learning linear functions, you are still at the beginning and are learning the basics or whether the topic is already more advanced.*
>
> Chinese student, 06SSH19: *Learning state means how enthusiastic you are while you are learning mathematics or which learning experiences you make and how much energy you invest. For example, whether you have slept well and have made a mental effort today.*

Consequently, German and Chinese (Shanghai) students showed pronounced differences with regard to item interpretation and associated experiences in classroom. German students more often referred to **teachers' goals** aiming at the students' understanding and to experiences in classroom in which the teacher **monitors** understanding (with exams, oral assessments) or uses **teaching practices** to eliminate misunderstandings of struggling students ("remedial activities", this was mentioned significantly more often by German students: $\chi^2(1)=9.33$, $p = .00$). In contrast, Chinese (Shanghai) students significantly more often referred to **goals** aiming at the students' attention (χ^2

(1) =15.56, p = .00) and to experiences in classroom in which the teacher **monitors** attention (e.g., by controlling their homework, χ^2 (1) =13.26, p = .00), or uses **teaching practices** that foster students' attention,

> German student, 01DSHA13: *(...) it means that our teacher wants to ensure that we understand the topics and subtopics of a mathematics lesson.*
>
> German student, 03SER24: *(...) if we don't understand, he [the teacher] explains it once more, or we work in learning groups. The students who have understood help other students/groups who have not understood.*
>
> Chinese student, 02SSH29*: (...) it means that the teacher wants to ensure that students are not distracted and listen carefully during instruction.*
>
> Chinese student, 02SSH11*: That means that during instruction some students listen attentively, some are in the classroom yet they are thinking of something else, some fall asleep. Then, the teacher wakes them up.*

Thus, our results support *Hypothesis 2* that interpretation differs strongly for statistically non-comparable items. These variations are linked to translation differences but also to different culture-specific definitions of support: advancing understanding in Germany, attention and effort in China (Shanghai). Further, regardless of group membership, CCCIs identified that students had difficulties with "showing interest" as it refers to the teacher's intention, which is hardly observable for students. Instead, students reported observable behaviours.

3.2.2 Item TS04: The teacher helps students with their learning

Statistically, Item TS04 was not comparable. For Item TS04 both the German and Chinese (Shanghai) version adhered to a literal translation of the term *learning.* Reviewing the transcripts, we noted some overlap but also variation in item interpretation and associated experiences. The majority of students from both groups (13 out of 14 students each) associated "helping with learning" with **teaching practices** that eliminate students' learning difficulties ("remedial activities") with a focus on solving mathematical tasks, particularly in China (Shanghai),

> German student, 08DSAS15*: (…) for me, it means that the teacher approaches students who may have problems solving a mathematical task, or who are stuck. Of course she helps us and explains it once more.*
>
> Chinese student, 04SSH26: *If we cannot solve a mathematical task, she [the teacher] will explain it to us step by step (…) until we understand everything.*

Besides similarities, CCCIs also identified variations in item interpretation. "Helping with learning" was additionally associated with experiences of **content-focused teaching.** However, German students significantly more often referred to the provision of additional learning opportunities (χ^2 (1)=18.12, p = .00), while Chinese (Shanghai) students significantly more often referred to being taught how to solve mathematical tasks with various methods (χ^2 (1)=12.13, p = .00),

> German student, 02DSHA23*: (…) if we get rules of thumb that helps us with our learning to a great extent (…) if something is explained, and just briefly said that it works that way, that doesn't help us with our learning a lot, but if we have to solve mathematical tasks, and think 'hey, how does that work?' then, with a rule of thumb (…), then it's much easier to remember, and consequently also [easier] to solve the task.*
>
> German student, 02DSHA23*: (…) or the use of graphs when the teacher explains something on the whiteboard. For instance, today, we learned about irrational numbers, integers, and decimals, we drew a circle around them, and then you know that the integers also appear in the decimals, because the circle always expands outwardly and, thus, the inner circle is always included.*
>
> Chinese student, 04CSSH26: *(…) well for us, the students in the 9th grade that means that the teacher constantly helps us to solve mathematical tasks and explains them during instruction.*

In contrast to the Chinese (Shanghai) students, German students additionally referred to relationship-oriented experiences, such as **teaching practices** that encourage learning with other students, **socio-emotional experiences** supporting the student-teacher relationship, and, even though the item explicitly refers to the teacher, **autonomous student learning** (these differences were significant: χ^2(1)=4.67, p = .03),

> German student, 01SHA19*: (…) helping with learning, also means to treat us with respect. I think that is also important for students, because you feel that you are being taken seriously, and that you are not just someone who simply doesn't understand.*

German student, 05DSMI26*: (…) often we have to explain it to each other during group work, so that we do not always depend on her [the teacher], because, if you do it at home, you do not have her either.*

German student, 07DSEL25*: (…) she [the teacher] explains so that we understand, but learning, memorizing something, that is something you do alone or with peers (…) me and my friend prepared the last math exam in the afternoon, during a phone call we checked all exercises and homework once more, everything that we had got wrong or that we did not understand, then we explained it to each other or came up with the solution together, that helps.*

To sum up, in China (Shanghai) "helping with learning" was understood as teaching the students how to solve mathematical tasks, and, in case they are struggling, by applying remedial activities and adaptive support. In Germany, students also often referred to remedial activities, yet, the interpretative pattern showed more variation. Thus, *Hypothesis 2* was supported, as we found interpretative variations for a statistically non-comparable item. Further, CCCIs identified an interpretative overlap with the previous item TS01 for Germany – both items were associated with remedial activities - which was also noted by some German students ("as I already said", "similar to the previous item").

3.2.3 Item TS05: The teacher continues teaching until the students understand

Statistically, Item TS05 was comparable. While the source version refers to "continues teaching", both the German and the Chinese (Shanghai) translation mean "continues *explaining*". A first glance at the interpretative patterns showed interpretative similarities. All 28 interviewed students associated "explaining until the students understand" with **teaching practices** to eliminate comprehension problems and to ensure understanding ("remedial activities", in China (Shanghai) with a focus on solving mathematical tasks),

German student, 05DSNI22: *(…) if the students have not understood something yet, the teacher explains it once more, and afterwards checks if every student has understood.*

Chinese student, 06SSH19: *That means, (…) sometimes students cannot understand what the teacher has explained, so the teacher will continue to explain until the students understand and can solve the mathematical tasks on their own.*

However, the students referred to different situational contexts. In Germany, students referred to experiences during instruction. They reported that the teacher either continues explaining for all or for individual students, or encourages other students to help their peers, while students in China (Shanghai) referred to individual support after instruction,

> German student, 05DSBA25: *If, for instance, the majority doesn't understand, she [the teacher] explains it once more in detail. If a small group of students doesn't understand, she explains once more specifically for the individual students on the whiteboard.*
>
> German student, 03SER24: *In case the student still doesn't understand (…) then another student is asked to explain, as students use different examples that are more suitable for students and young people.*
>
> Chinese student, 01SSH24: *In case we do not have enough time during instruction, she [the teacher] asks the students who don't understand to come to the teachers' room after instruction*.

If German students mentioned experiences after instruction, they referred to external help,

> German student, 07DSEL25: *If we do not understand you can also ask different people, parents if they are able to help you, or peers. And there are also many YouTube videos that explain it.*

Further, the in-depth probing of what "until students understand" means identified substantive differences. With regard to **time management,** German students significantly more often reported that *until* means that the teacher more or less flexibly takes the time that is needed during instruction ($\chi^2(1)=7.64$, $p = .01$). Students from Shanghai, however, understood *until* as one to maximum ten minutes (depending on the difficulty of the mathematical task) and significantly more often reported that the time for additional explanations during instruction is fixed (e.g., due to the curriculum or the pace of instruction; $\chi^2(1)=9.33$, $p = .00$).

> German student, 07DSGO25: *She [the teacher] explains again and again, until you understand. Sometimes you simply do not understand, but she does not say, enough, we can ask her again.*
>
> Chinese student, 02SSH29: *Approximately 3 to 5 minutes. It [additional explanation] doesn't need much time, otherwise instruction would be hindered.*

We also found differences with regard to the **reference group.** While German students significantly more often associated *students* with the majority of students, *students* was

significantly more often associated with high-achieving students in China (Shanghai) ($\chi^2(1)=6.09$, $p = .01$, $\chi^2(1)=4.68$, $p = .03$ respectively),

> German student, 07DSGO25: *(...) she continues explaining to all students (...). Until the majority understands.*
>
> Chinese student, 05SSH15: *No, she doesn't continue to explain for all students. Difficult mathematical tasks are not obligatory for low-achieving students. For the smart high-achieving students, she would continue to explain.*
>
> Chinese student, 04SSH26: *When high-achieving students have understood everything, it does not matter very much if the low-achieving students have understood or not. The teacher always explains things. However, some low achieving students don't listen (...) consequently they do not have any questions. We, the high-achieving students, normally have many questions, the teacher explains thoroughly until we have understood everything.*

Hence, students from Germany and China (Shanghai) agreed with regard to the semantic meaning. Yet, depending on the context, different criteria define when "until students understand" is accomplished. As interpretations differed for a statically comparable item, *Hypothesis 1* was not supported. Moreover, CCCIs identified interpretative similarities between Item TS05 and the previous items, in Germany particularly with Item TS01 and in China (Shanghai) with Item TS04.

4 Discussion

The aim of this study was to investigate the impact of culture on item interpretation and to identify sources of interpretative variations. To this effect, we conducted cognitive interviews with students from Germany and China (Shanghai) and comparatively analysed their interpretations of PISA student support items.

4.1 Culture-specific interpretative variations: sources and implications

At least three things can be learned from our study: First, culture can have an impact on the cognitive processing of survey items and, consequently, can compromise data comparability. Secondly, much can be learned by integrating qualitative and quantitative methods. For statistically

non-comparable items, CCCIs identified distinctive variations in item interpretation between the interviewed students from Germany and China (Shanghai) (supporting *Hypothesis 1*). However, we also found some differences in interpretative patterns for the item that was not flagged to be biased in our statistical analysis (*Hypothesis 2* was not confirmed). And lastly, CCCIs proved to be a compelling means of identifying sources of variation. Item interpretation was shaped by various culture-specific factors, which we discuss in the following.

4.1.1 Instrument characteristics

The first source of interpretative variations are instrument characteristics, such as translation differences. Cross-cultural studies including PISA usually implement rigorous translation and verification procedures to ensure the equivalence of the source questionnaire and the respective national versions (OECD, 2014). However, the equivalence between the various national and linguistic questionnaires is not checked. National questionnaires are developed by native speakers with a specific cultural background, which may manifest in translation, and lead to a shift in item meaning to match the cultural context and consequently trigger off culture-congruent interpretations. For one item, CCCIs identified culture-specific translations of the term *learning* (linked to differences in teaching traditions, see Section 4.1.2, and the definition of student support, see Section 4.1.3), which led to distinct interpretative variations between the interviewed students from Germany and China (Shanghai). Strikingly, these different translations have been administered in PISA since 2000, producing data with limited comparability over years. As translations are verified by native speakers with a similar cultural background, culture-specific variations may often remain undetected.

Further, item complexity and content can trigger off divergent interpretations. Students had difficulties with complex items involving more than one concept (e.g., showing interest in students'

learning). This is in line with findings on a national level for classroom management and cognitive activation (see Lenske, 2016). Complex wordings can be a serious threat to the cognitive validity, as students tend to reduce the complexity which can shift the item meaning (Lenske, 2016). Similarly, students focused on overt behaviours, yet they barely referred to abstract concepts (e.g., the teacher's interest). As these are not directly observable for students, they have to rely on indirect behavioural indicators, which students might not interpret correctly (Fauth, Göllner, Lenske, Praetorius, & Wagner, 2020). Moreover, students from Germany and China (Shanghai) showed variations with regard to the interpretation of ambiguous terms (e.g., teaching *until* students understand), which were linked to preferred instructional approaches and goals, for instance with regard to time management and targeted student group. Further, in a cross-cultural context, items involving the broad term *learning* seem to show a low comparability (see also Fischer et al., 2019). Similarly, previous studies have pointed towards the advantage of shorter, simpler, and unambiguous items in increasing intercultural comparability (Harkness, van de Vijver, & Mohler, 2003).

Finally, scale characteristics can have a bearing on interpretation. Particularly for Germany, the wording of the individual items was rather similar, which may have led to an interpretative overlap (e.g., all items were associated with remedial activities). On the other hand, compared to Chinese students, German students may be more likely to endorse item ordering effects (Haberstroh, Oyserman, Schwarz, Kühnen, & Ji, 2002). Further, the scale did not fully match the teaching reality in China (Shanghai). The interviewed students often referred to support after instruction, yet the question, item, and answering categories specifically refer to the actual lessons. Thus, support might be underestimated in China (Shanghai). Consequently, to enhance cross-cultural comparability, culture-specific interpretative variations based on instrument characteristics have to

be considered during instrument development, and translation and culture should be disentangled to some degree.

4.1.2 Mathematics education in different cultural traditions

The cultural context underlying students' education can have a bearing on item interpretation. Context information and experiences that are applied to make sense of the item content (*information retrieval stage*) are highly culture and situation depended and can shift the item meaning. For instance, in order to pursue secondary education after nine years of compulsory education, students in Shanghai have to take a public examination (*Zhongkao*) at the age of 15 (OECD, 2016). This may be one reason why Shanghai students almost exclusively referred to experiences in classroom focusing on solving math problems. The preparation for the examination might also lead to a higher need of support compared to "regular" teaching phases.

Moreover, culture-specific instructional goals and approaches can shape item interpretation. Leung (2001) characterises mathematics instruction in Western versus East-Asian countries along six dichotomies, which can be linked to the interpretative variations we found in our study. Accordingly, mathematics instruction in many Western countries is *process* oriented (e.g., how mathematical knowledge is arrived at) and student-centred with a focus on *individualized learning*. The teacher is seen as a facilitator of learning, thus, *competence in pedagogy* is important (e.g., the teacher teaches the students to be autonomous or how to acquire knowledge from other sources). *Pleasurable* (e.g., interesting tasks such as group work) and *meaningful learning* (e.g., thorough understanding) are seen as the heart of effective student learning, while *intrinsic* factors are important to motivate students to learn. Mathematics instruction in most East-Asian countries is characterized as being *product* oriented, with the aim of getting the body of knowledge from the teacher to the students in a teacher-directed *whole class* setting. The teacher is seen as a source of knowledge, thus, *competence in subject matter* is important. *Studying hard* (e.g., effort and

attention) and *root learning* (e.g., practicing and memorizing) are seen as the heart of student learning, while *extrinsic* factors are important to motivate students to learn. Hence, teaching can be described as cultural activity (Stigler & Hiebert, 1999), which can challenge the comparative validity of student support items. The translation variations we found for the term *learning* may be linked to different traditions, whereupon effective learning in Germany depends on the quality of the learning progress and the students' ability, while effort and hard work are important for success in China (Leung, 2006). The latter corresponds to the Confucian presumption that differences in ability do not inhibit ones educability, everyone is educable as long as enough effort is invested (Leung, Park, Shimizu, & Xu, 2015). Further, based on education traditions, the targeted outcomes may differ across education systems as well as the preferred practices that are used to satisfy goals. Across all items, the interviewed Chinese students often referred to practices fostering their attention, effort, and learning how to solve mathematical tasks, while the process of eliminating misunderstanding was frequently reported in Germany. Both practices are likely to have a predictive validity on students' outcomes, however, the targeted outcome seems to be the ability to solve mathematical tasks in China (Shanghai) but mathematical understanding in Germany.

4.1.3 Definition of student support

Another source of interpretative variations are culture-specific construct definitions. The student support items were often associated with support of competence. Yet, while German students referred to a flexible adjustment of the instructional pace according to all students' needs and remedial activities to support struggling students, competence support in China (Shanghai) was additionally referred to as fostering understanding in high-achieving students. This is in line with findings that show a higher perceived student support for disadvantaged students in Germany respectively for advantaged students in China (Shanghai) (OECD, 2019, 2020). Students from China (Shanghai) noted, however, that competence support extends beyond the boundaries of the

classroom, and after-school lessons are provided for struggling students but also for enrichment. The focus on competence support may be due to the fact that the items target practices that support students' learning. Nevertheless, some German students referred to socio-emotional experiences, such as learning in groups or pairs, respect, and a teacher that shares private stories and has fun during instruction. This difference might be linked to the fact that for Western teachers, it is important that the content they teach is meaningful and interesting for students, while East-Asian teachers tend to adhere to the syllabus and textbooks (Leung, 2006).

Students from China (Shanghai) additionally referred to practices that foster academic productiveness, namely keeping students attentive, on task, and to avoid a waste of time. According to the theoretical conceptualization of teaching quality, these are sub-dimensions of classroom management and not student support (Klieme et al., 2009). This can have severe consequences for the comparability of student ratings as the interpretations of the interviewed Chinese (Shanghai) students do not correspond to the researchers' intentions. Three explanations are conceivable: Besides translation variations, video studies suggest that classroom management is more important than student support in mathematics lessons in Shanghai (Opfer et al., in press). Moreover, the interviewed Chinese students may have different perceptions of support, whereupon supporting the students' effort is more important than supporting their understanding or socio-emotional experiences. Interpretative variations can be an indicator that instruments based on theories and models developed in a certain context are not necessarily suitable in others (Lenske & Praetorius, 2020). The PISA student support scale was developed based on work by Carroll (1989) and Stringfield and Slavin (1992) based on teaching in western countries. In order to improve cross-cultural measurement, future research should more closely investigate the conceptualization of student support in China (Shanghai) and adapt the items to the Chinese context.

4.2 Strengths and limitations

It is important to note some limitations of this research, which may influence the interpretation of the findings and provide reasonable directions for future studies.

To study the impact of culture on item interpretation, we selected two cultural groups for which data comparability has been demonstrated to be particularly limited, and a construct that is likely shaped by cultural and personal perceptions. Further, we evaluated items that have been administered in PISA since 2000. Yet, bias is always a function of an instrument applied in a specific context (Benítez, van de Vijver, & Padilla, 2019). Consequently, further research should be conducted to investigate if our results are generalizable for additional constructs, instruments, and education systems. Moreover, we aimed at evaluating the cognitive validity of interpretations of mainstream students and our study should be extended to different subgroups within education systems, for which interpretative patterns might vary (e.g., students with migration background, different school types).

CCCI is indisputably one of the most suitable methods to identify differences and similarities in the survey response process. While the cross-cultural context provides a unique opportunity, it also goes hand in hand with several challenges. First, the Chinese (Shanghai) and German sample for the CCCIs' slightly differed with regard to the average last math grade and number of books at home. This might impact the results. Further, it is nearly impossible to establish an absolutely comparable interviewing situation when different cultures are involved. To standardize the interviewing situation and eliminate interviewer effects, the same German interviewer conducted the probing in Germany and China (Shanghai). This procedure, however, entails the risk of effects caused by the presence of a foreigner together with an interpreter in China (Shanghai). Further, cross-cultural studies involve participants and researchers with different native languages. To avoid

a third working language, both the German and Chinese (Shanghai) interview data was transcribed into German. To ensure a high degree of accurate translations, the translator was not only a Chinese native speaker, but also a math expert. This might have led to a slight modification of the students' responses during translation, for instance through the use of more technical notions. Further, the initial coding system was developed based on German data and revised after applying it to the Chinese (Shanghai) interview data. While this procedure helped us to identify if codes are assigned for both Germany and China (Shanghai) or are country-specific, a different code development procedure might have led to a different final coding system (e.g., a joint analysis of German and Chinese data). Lastly, the researchers' perspectives are shaped by their own cultural background, which can have a bearing on several phases of the study. To avoid a unilateral perspective, the German and Chinese research team collaborated closely during all phases of the study.

5 Conclusions

Regardless of the importance of student support on learning, our results highlight that measuring student support in a cross-cultural context is a challenging task. We found variations in item interpretation for statistically non-comparable but also for a comparable item, which were based on belonging to one specific group, with a specific language, in a specific cultural context. Hence, the operationalization and measurement of educational constructs should not only rely on statistical criteria, but also be based on a shared conceptual understanding. Thus, we strongly recommend the combination of statistical tests of data comparability with cognitive methods studying how respondents mentally process and respond to survey items (Meitinger, 2017). We urge future research to evaluate the impact of culture on item interpretation for additional constructs and instruments in order to avoid false conclusions.

References

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*, 495–508. https://doi.org/10.1080/10705511.2014.919210

Behr, D., Braun, M., Kaczmirek, L., & Bandilla, W. (2014). Item comparability in cross-national surveys: results from asking probing questions in cross-national web surveys about attitudes towards civil disobedience. *Quality & Quantity, 48*, 127–148. https://doi.org/10.1007/s11135-012-9754-8

Benítez, I., & Padilla, J.-L. (2014). Analysis of nonequivalent assessments across different linguistic groups using a mixed methods approach: understanding the causes of differential item functioning by cognitive interviewing. *Journal of Mixed Methods Research*, *8*, 52–68. https://doi.org/10.1177/1558689813488245

Benítez, I., van de Vijver, F.J.R., & Padilla, J.-L. (2019). A mixed methods approach to the analysis of bias in cross-cultural studies. *Sociological Methods & Research*, *8*. https://doi.org/10.1177/0049124119852390

Carroll, J. B. (1989). The Carroll Model. A 25-year retrospective and prospective view. *Educational Researcher, 18,* 26–31.

Clarke, D., Keitel, C., & Shimizu, Y. (2006). *Mathematics classrooms in twelve countries - The insider's perspective*. Rotterdam: Sense Publishers.

Crowne, D. P., & Marlowe, D. (1964). *The approval motive: studies in evaluative dependence.* Hoboken, NJ: John Wiley & Sons.

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.* Paris: OECD Publishing.

Dorfner, T., Förtsch, C., & Neuhaus, B. J. (2018). Effects of three basic dimensions of instructional quality on students' situational interest in sixth-grade biology instruction. *Learning and Instruction, 56,* 42-53. https://doi.org/10.1016/j.learninstruc.2018.03.001

Fauth, B., Decristan, J., Decker, A.-T., Büttner, G., Hardy, I., Klieme, E., & Kunter, M. (2019). The effects of teacher competence on student outcomes in elementary science education: the

mediating role of teaching quality. *Teaching and Teacher Education, 86*. https://doi.org/10.1016/j.tate.2019.102882

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: dimensions and prediction of student outcomes. *Learning and Instruction, 29,* 1-9. https://doi.org/10.1016/j.learninstruc.2013.07.001

Fauth, B., Göllner, R., Lenske, G., Praetorius, A.-K., & Wagner, W. (2020). Who sees what? Conceptual considerations on the measurement of teaching quality from different perspectives. *Zeitschrift für Pädagogik*, *66*, 138–155.

Fischer, J., He, J., & Klieme, E. (2020). The structure of teaching practices across countries: a combination of factor analysis and network analysis. *Studies in Educational Evaluation, 65,* https://doi.org/10.1016/j.stueduc.2020.100861

Fischer, J., Praetorius, A.-K., & Klieme, E. (2019). The impact of linguistic similarity on cross-cultural comparability of students' perceptions of teaching quality. *Educational Assessment, Evaluation and Accountability*, *31*, 201–220. https://doi.org/10.1007/s11092-019-09295-7

Fitzgerald, R., Widdop, S., Gray, M., & Collins, D. (2011). Identifying sources of error in cross-national questionnaires: application of an error source typology to cognitive interview data. *Journal of Official Statistics, 27,* 569-599.

Grisay, A., de Jong, J. H. A. L., Gebhardt, E., Breezier, A., & Halleux-Monseur, B. (2007). Translation equivalence across PISA countries. *Journal of Applied Measurement, 8,* 249–266.

Haberstroh, S., Oyserman, D., Schwarz, N., Kühnen, U., & Ji, L.-J. (2002). Is the interdependent self more sensitive to question context than the independent self? Self-construal and the observation of conversational norms. *Journal of Experimental Social Psychology*, *38*, 323-329. https://doi.org/10.1006/jesp.2001.1513

Harkness, J. A., van de Vijver, F.J.R., & Mohler, P.P. (2003). *Cross-Cultural Survey Methods*. Hoboken, NJ: John Wiley & Sons.

Harzing, A.-W. (2006). Response styles in cross-national survey research: a 26-country study. *International Journal of Cross Cultural Management*, *6*, 243–266. https://doi.org/10.1177/1470595806066332

Hattie, J. (2009). *Visible learning*. London: Routledge.

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. https://doi.org/10.1177/0022022116687395

He, J., & Kubacka, K. (2015). Data comparability in the teaching and learning international survey (TALIS) 2008 and 2013. *OECD Education Working Papers*. Paris: OECD Publishing.

Hofstede, G. (2001). *Culture's consequences: comparing values, behaviors, institutions, and organizations across nations.* Thousand Oaks, CA: Sage Publications.

Hopfenbeck, T. N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J., & Baird, J.-A. (2017). Lessons learned from PISA: a systematic review of peer-reviewed articles on the Programme for International Student Assessment. *Scandinavian Journal of Educational Research, 62*, 333-353. https://doi.org/10.1080/00313831.2016.1258726

Karabenick, S. A., Woolley, M. E., Friedel, J. M., Ammon, B. V., Blazevski, J., Bonney, C. R., de Groot, E., Gilbert, M.C., Musu, L., Kempler, T.M., & Kelly, K. L. (2007). Cognitive processing of self-report items in educational research: do they think what we mean? *Educational Psychologist*, *42*, 139–151. https://doi.org/10.1080/00461520701416231

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.* Münster: Waxman.

Koskey, K. L.K., Karabenick, S. A., Woolley, M. E., Bonney, C. R., & Dever, B. V. (2010). Cognitive validity of students' self-reports of classroom mastery goal structure: what students are thinking and why it matters. *Contemporary Educational Psychology, 35,* 254-263. https://doi.org/10.1016/j.cedpsych.2010.05.004

Kuckartz, U. (2018). *Qualitative Inhaltsanalyse: Methoden, Praxis, Computerunterstützung [Qualitative content analysis: methods, practices, computer-based analysis]*. Weinheim, Basel: Beltz Juventa.

Lenske, G. (2016). *Schülerfeedback in der Grundschule [Student feedback in elementary school]*. Münster: Waxmann.

Lenske, G., & Praetorius, A.-K. (2020). Schülerfeedback – was steckt hinter dem Kreuz auf dem Fragebogen? [Student feedback – what is the basis for choosing an answer on a

questionnaire?] In M. Clausen & K. Göbel (Eds.). *Unterrichtsrückmeldungen durch Schüler*innen.* Empirische Pädagogik, 34, 11-29.

Leung, F.K.S. (2001). In search of an East Asian identity in mathematics education. *Educational Studies in Mathematics*, *47*, 35–51. https://doi.org/10.1023/A:1017936429620

Leung, F.K.S. (2006). Mathematics education in East Asia and the West: Does culture matter? In F. K. Leung, K.-D. Graf, & F. J. Lopez-Real (Eds.), *Mathematics education in different cultural traditions - a comparative study of East Asia and the West.* Boston, MA: Springer International Publishing.

Leung, F. K. S., Park, K., Shimizu, Y., & Xu, B. (2015). Mathematics education in East Asia. In S. J. Cho (Ed.), *The Proceedings of the 12th international Congress on Mathematical Education*. Springer International Publishing. https://doi.org/10.1007/978-3-319-12688-3_11

Lietz, P. (2017). Design, development and implementation of contextual questionnaires. In P. Lietz, J. C. Cresswell, K. F. Rust, & R. J. Adams (Eds.), *Implementation of large-scale education assessments*. Chichester, West Sussex, United Kingdom: John Wiley & Sons.

Mesiti, C., & Clarke, D. (2017). The international lexicon project: giving a name to what we do. In R. Seah, M. Horne, J. Ocean, & C. Orellana (Eds.), *Proceedings of the Mathematical Association of Victoria annual conference*. Mathematical Association of Victoria.

Miller, K., Fitzgerald, R., Padilla, J.-L., Willson, S., Widdop, S., Caspar, R., Dimov, M., Gray, M., Nunes, C., Prüfer, P., Schöbi, N., Schoua-Glusberg, A., & Willis, G.B. (2011). Design and analysis of cognitive interviews for comparative multinational testing. *Field Methods*, *23*, 379–396. https://doi.org/10.1177/1525822X11414802

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide. Eighth 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. 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. (2014). *PISA 2012 technical report*. Paris: OECD Publishing.

OECD. (2016). *Education in China: a snapshot*. Paris: OECD Publishing.

OECD. (2019, 2020). *PISA 2018 technical report*. Paris: OECD Publishing.

Opfer, V. D., Bell, C. A., Klieme, E., McCaffrey, D. F., Schweig, J. D. & Stecher, B. M. (in press). Understanding and measuring mathematics teaching practice in eight countries and economies from four continents. In OECD (Eds.), *Global Teaching InSights: A video study of teaching*. OECD Publishing. https://doi.org/10.1787/20d6f36b-en

Paulhus, D. L. (1991). Measurement and control of response biases. In J. Robinson, P.R. Shaver, & L.R. Wrigthsman (Eds.), *Measures of personality and social psychological attitudes*. San Diego, CA: Academic Press. https://doi.org/10.1016/B978-0-12-590241-0.50006-X

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.

Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018b). Generic dimensions of teaching quality: the German framework of Three Basic Dimensions. *ZDM Mathematics Education 50*, 407–426. https://doi.org/10.1007/s11858-018-0918-4

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. https://doi.org/10.1016/j.learninstruc.2013.12.002

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.* https://doi.org/10.3389/fpsyg.2016.00110

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

Schwarz, N. (2007). Cognitive aspects of survey methodology. *Applied Cognitive Psychology, 21,* 277–287. https://doi.org/10.1002/acp.1340

Schwarz, N., Oyserman, D., & Peytcheva, E. (2010). Cognition, communication, and culture: implications for the survey response process. In J. A. Harkness (Ed.), *Survey methods in multinational, multiregional, and multicultural contexts*. Hoboken, N.J.: John Wiley & Sons. https://doi.org/10.1002/9780470609927.ch10

Stigler, J.W., & Hiebert, J. (1999). *The teaching gap: best ideas from the world's teachers for improving education in the classroom.* New York: The Free Press.

Stringfield, S. C., & Slavin, R. E. (1992). A hierarchical longitudinal model for elementary school effects. In B.P.M. Creemers & G. J. Reezigt (Eds.), *Evaluation of educational effectiveness*. Groningen: ICO.

Tourangeau, R. (1984). Cognitive science and survey methods: a cognitive perspective. In T. Jabine, M. Straf, J.Tanur, & R. Tourangeau (Eds.), *Cognitive aspects of survey design: building a bridge between disciplines*. Washington, DC: National Academy Press.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511819322

Uskul, A. K., & Oyserman, D. (2006). Question comprehension and response: Implications of individualism and collectivism. In B. Mannix, M. Neale, & Y. Chen (Eds.), *Research on managing groups and teams: national culture & groups*. Amsterdam, Boston: Elsevier Science Press.

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

Van de Vijver, F. J.R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research.* Thousand Oaks, CA: Sage.

Van Herk, H., Poortinga, Y. H., & Verhallen, T. M. M. (2004). Response styles in rating scales: evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology*, *35*, 346–360. https://doi.org/10.1177/0022022104264126

Varnum, M. E. W., Grossmann, I., Kitayama, S., & Nisbett, R. E. (2010). The Origin of cultural differences in cognition: the social orientation hypothesis. *Current Directions in Psychological Science, 19*, 9–13. https://doi.org/10.1177/0963721409359301

VERBI Software (2018). *MAXQDA 2018.2 [computer software]*. VERBI Software. Retrieved from maxqda.com.

Vieluf, S., Kaplan, D., Klieme, E., & Bayer, S. (2012). *Teaching practices and pedagogical innovation: evidence from TALIS.* Paris: OECD Publishing.

Vieluf, S., Kunter, M., & van de Vijver, F. J.R. (2013). Teacher self-efficacy in cross-national perspective. *Teaching and Teacher Education, 35,* 92-103. https://doi.org/10.1016/j.tate.2013.05.006

Willis, G. B. (2005). *Cognitive interviewing: a tool for improving questionnaire design*. Thousand Oaks, CA: Sage.

Willis, G. B., & Miller, K. (2011). Cross-cultural cognitive interviewing: seeking comparability and enhancing understanding. *Field Methods*, *23*, 331–341. https://doi.org/10.1177/1525822X11416092

Wisniewski, B., Zierer, K., Dresel, M., & Daumiller, M. (2020). Obtaining secondary students' perceptions of instructional quality: two-level structure and measurement invariance. *Learning and Instruction, 66*. https://doi.org/10.1016/j.learninstruc.2020.101303

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. https://doi.org/10.1007/s12564-016-9455-4

Appendix C: Manuscript 3

Fischer, J., He, J., & Klieme, E. (2020). The structure of teaching practices across countries: A combination of factor analysis and network analysis. *Studies in Educational Evaluation, 65*. https://doi.org/10.1016/j.stueduc.2020.100861

Studies in Educational Evaluation 65 (2020) 100861

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/locate/stueduc

The structure of teaching practices across countries: A combination of factor analysis and network analysis

Jessica Fischer[a,1,*], Jia He[a,b,1], Eckhard Klieme[a]

[a] DIPF | Leibniz Institute for Research and Information in Education, Frankfurt, Germany
[b] Tilburg University, The Netherlands

ARTICLE INFO

Keywords:
Teaching practices
Classroom assessment
Factor analysis
Network analysis
PISA
Cross-cultural

ABSTRACT

Teaching practices are pivotal for student learning. Due to pedagogical traditions and national cultures, the structure of teaching practices may differ across countries. This study investigates the structure of teaching practices across 12 countries grouped into four major linguistic/cultural clusters. First, factor analysis is applied to investigate if the theoretical distinction between teacher-directed and student-centred practices is generalizable across countries. Then, network analysis is used to explore how individual classroom assessment practices relate to either teacher-directed or student-centred practices. Main findings include that: (1) teacher-directed and student-centred practices are two distinct factors across countries; (2) the overall structure and connectivity of teaching practices differs across countries, with smaller differences within linguistic/cultural clusters; and (3) assessment practices with the aim to structure and guide learning strongly relate to teacher-directed practices, whereas assessment practices with the aim to individualize instruction more relate to student-centred practices. We discuss the global patterning and implications.

1. Introduction

Across the world, teacher's instructional practice has been recognized to be one of the most important factors influencing student learning outcomes (Hattie, 2009). Teaching practices can be seen as a major part of classroom instruction and in contrast to other factors relevant for student learning (e.g., the student's socio-economic background) they are more readily modifiable and, thus, can be subjected to targeted interventions (Vieluf, Kaplan, Klieme, & Bayer, 2012). In the last decades, international research often discussed two approaches to teaching, deemed opposite to each other, based on philosophies of education: teacher-directed and student-centred teaching practices (Tobias & Duffy, 2010). However, it has been argued that these theoretical conceptualizations do not account for the complex nature of teaching practices. First, there is no single best way of teaching; instead teachers are required to combine various strategies depending on the context, class, and students (Echazarra, Salinas, Méndez, Denis, & Rech, 2016). Research to identify how various practices relate to each other and what is the most beneficial mix is still scarce. Second, teaching can be regarded as cultural activity and is not generalizable across countries (Stigler & Hiebert, 1999). National culture and pedagogical tradition interactively influence approaches to teaching, leading to differences in frequency and combination of teaching practices, and consequently challenging the assumption of a universal structure of teaching practices. Additionally, the international debate recently considers classroom assessment (e.g., providing feedback) as part of instructional practices that teachers implement in the classroom (OECD, 2013). Yet, it remains unclear how classroom assessment relates to teacher-directed and student-centred practices. In order to tailor targeted interventions to promote high-quality teaching in a culturally sensitive way, a comprehensive understanding of the structure and co-occurrence of teaching practices across countries is indispensable.

In this study, we aim to shed light on the structure and co-occurrence of teaching practices across countries in two steps. First, we check if the theoretical distinction between teacher-directed and student-centred practices is empirically supported across countries. Secondly, we investigate how classroom assessment practices (which were rarely simultaneously tested with teacher-directed and student-centred practices in empirical studies) integrate into the broader framework of teaching practices with an exploratory approach. More precisely, we investigate if individual classroom assessment practices differently relate to either teacher-directed or student-centred practices. Given that high-quality teaching requires teachers to combine diverse practices to foster student learning, we propose to consider direct relationships

* Corresponding author.
E-mail address: jessica.fischer@dipf.de (J. Fischer).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.stueduc.2020.100861
Received 5 July 2019; Received in revised form 24 January 2020; Accepted 21 February 2020

between individual practices beyond focusing on the shared underlying factors. Thus, we complement conventional factor analysis with psychological network analysis. Network analysis models direct interactions among individual practices and helps visualize the "ecosystem" (e.g., conditional co-occurrence) of teaching practices. It helps us 1) to illustrate the conditional co-occurrence of practices and compare the patterning across countries, 2) to account for the interdependency without reducing the related practices to a single construct score, as has been done in studies on teaching practices (see e.g., OECD, 2019), and 3) to provide a foundation to further explore overarching teaching quality dimensions in the future.

In the following, we first review the framework of teaching practices and recent developments, and highlight the importance of a cross-cultural perspective on teaching practices. Thereafter, we address the challenges of measuring and analysing teaching practices across countries.

1.1. Teacher-directed versus student-centred teaching practices

In the 20th century, educational theories have undergone significant developments. Influenced by the *behaviourism* in the United States (Carroll, 1963), and the German schools *Reform pedagogy* (e.g., by Peter Petersen) and *Gestalt psychology* (Duncker & Lees, 1945) in Western Europe, instructionist and constructivist theories of learning have emerged. Rooted in Western countries, both frameworks are increasingly influential outside North America and Western Europe. Instructionists such as Rosenshine (1976) characterize a traditional and teacher-directed approach to instruction with an information-processing view of learning. In contrast, constructivism, based on work of Vygotsky (1978), Dewey (1929), and Piaget (1952), promotes an alternative approach with the focus on the learner and learning context (Tobias & Duffy, 2010). These two dominant frameworks inspire approaches to designs of instruction to date, yielding the development and application of different teaching practices.

Direct instruction advocates the use of *teacher-directed* practices that aim to provide a well-structured and effective learning environment (Caro, Lenkeit, & Kyriakides, 2016). The teacher is the transmitter of knowledge and controls learning processes in the classroom. Besides delivering information, the teacher plans lessons in advance and structures the presentation of ideas in class. Guided by the teacher, students can participate during instruction, for instance, through answering the teacher's questions, posing own questions to the teacher, or reproducing received information (Mostafa, Echazarra, & Guillo, 2018). The advantage of teacher-directed practices is the emphasis on well-structured lessons, wider subject coverage, and a better preparation for standardized tests (Ormrod, 2012). Yet, the rather passive role of students can lead to a decline in motivation and positive attitudes towards the subject (Echazarra et al., 2016).

Student-centred teaching practices based on constructivism foster students' active engagement in their learning processes and promote a self-directed construction of knowledge. This can be facilitated by assigning activities that involve students in planning classroom activities, promoting discussions among students themselves and with the teacher, or by creating cooperative learning environments (e.g., small group work) while taking individual students' needs into account (e.g., achievement levels). The role of the teacher is to support and guide the learning processes. Student-centred practices are supposed to foster communication skills and collaborations, encourage students to direct their learning, and develop interest in a subject. Yet, student-centred practices are harder to implement and are often criticized to lack guidance for the learner, overtaxing working memory (Tobias & Duffy, 2010), and risking the development of incorrect knowledge (Mostafa et al., 2018).

In reality, teaching is often a combination of diverse practices (Klieme, 2020). In line with this reasoning, educational effectiveness research criticises this theoretical distinction to be insufficient to fully benefit student learning. Instead, the complementary application of both teacher-directed and student-centred practices is often seen to be more effective (Vieluf et al., 2012). Thus, the co-occurrence of teaching practices should be considered when conceptualizing and analysing teaching practices.

1.2. Classroom assessment practices

Teacher-directed and student-centred practices stem from learning theories dating back to decades ago (Richardson, 2003). More recently, classroom assessment, as an additional element of teaching practices, has garnered much attention (OECD, 2013) and is considered one of the most powerful teaching practice for quality management and the improvement of student learning outcomes (Klieme, 2020). For instance, the Program for International Student Assessment (PISA), a triennial large-scale assessment of 15-year-old students in dozens of countries, operationalizes classroom assessment as a specific dimension of teaching practices (in addition to teacher-directed and student-centred practices) (OECD, 2014).

Classroom assessment practices are used to evaluate students' knowledge and progress (Coombs, DeLuca, LaPointe-McEwan, & Chalas, 2018). Depending on the standardization and purpose of the assessment, teachers possess a repertoire of tools to gather evidence about their students' progress and ideas (Harlen, 2007; Kippers, Wolterinck, Schildkamp, Poortman, & Visscher, 2018). These classroom assessment practices can serve the purpose of *summarizing* the achievement of students or the *formative* purpose of improving teaching and learning on an ongoing assessment basis (e.g., discussion or oral examination) (Black & Wiliam, 2009). Drawing on Ramaprasad's theory (1983), formative assessment includes three steps: 1) identifying the current learning state, 2) establishing learning goals, and lastly 3) defining the steps that are needed to achieve the learning goals. An integral part of formative assessment is feedback. Astin et al. (1996) suggest that assessment is most effective when diverse methods are implemented complementary. Especially formative assessment combined with feedback has been shown to be a powerful tool to improve student achievement and motivation (Harlen & Deakin-Crick, 2002; Hattie, 2009).

Echazarra et al. (2016) placed classroom assessment between both traditional (teacher-directed) and modern (student-centred) ends of a teaching practice scale, yet there is rarely empirical evidence supporting this classification. Moreover, as assessment practices have to be applied by teachers with either teacher-directed or student-centred approaches in order to identify the students' learning state and progress, the question remains if and how different assessment practices are incorporated into different approaches to teaching.

1.3. Teaching practices across countries

Both the constructivist and instructionist framework were mainly developed and empirically tested in Western countries, and they might not be easily transferable to other cultures. While policy-makers across the world have the consensus on the importance of promoting high-quality teaching, they may have a different understanding of the structure of teaching practices and the notions of good practices. Praetorius and colleagues (2018) surveyed educational researchers from different countries regarding what constitutes good practices in their respective country, and they found substantial cross-country differences with regard to the categorization of good practices depending on pedagogical traditions and national cultures. For instance, practices promoting deep thinking, students' autonomy, and adaptive teaching were especially important in South-American countries, whereas East-Asian countries mostly valued practices ensuring well-structured lessons and independent thinking. German researchers defined feedback, addressing student errors, orderly managing the class, cognitive activation, and social-emotional support as important practices in their

country. Furthermore, the effects of teaching practices on learning outcomes may be moderated by differences in educational systems or economic and cultural factors. For instance, Fuller and Clarke (1994) argued that student-centred practices promoting an active engagement of students during instruction are incompatible with strong hierarchical structures in countries valuing power distance. Likewise, McCormick and Alavi (2004) postulated that practices promoting teachers' critical reflection and inquiry might be less effective in collectivist countries, where criticism is communicated more indirectly than in Western countries. Consequently, the prevalent approaches to instruction and co-occurrence of teaching practices are likely to differ across countries. In a similar vein, cross-cultural research reported different frequencies of teacher-directed, student-centred, and classroom assessment practices across countries, yielding different teaching practice profiles (see e.g., Echazarra et al., 2016).

It is important to emphasize, that teaching practice is difficult to generalize; instead it can be described as "cultural activity" (see Stigler & Hiebert, 1999) as it exhibits vast cross-cultural differences, not only in quantity, but also in quality, processes, and effectiveness. Thus, it is vital to consider context-specificity when measuring and analysing teaching practices cross-culturally (Vieluf et al., 2012).

1.4. Measuring and analysing teaching practices across countries

The dynamic model of educational effectiveness by Creemers and Kyriakides (2006) proposes to refine the measurement of teaching and learning constructs along multiple dimensions including frequency (i.e., the quantity that an activity is present in a system, school, or classroom), focus (i.e., the specificity and purpose of an activity), stage (i.e., the phase of an activity, with the assumption that the activity needs to take place for a long period of time to accumulate effects on student learning), quality (i.e., properties of the activity and its optimal use), and differentiation (i.e., the extent to which the activity is implemented for and has impact on all subjects in the same way). These measurement dimensions capture not only quantity but also quality and processes. Methodologically, teaching practices can be assessed with self-reports in surveys (from teachers and/or students) (e.g., OECD, 2015) and behavioural coding in video studies (e.g., Jacobs, Hollingsworth, & Givvin, 2007). In large-scale educational assessment, where many countries are compared (quantitatively), survey-based measurement is more frequently applied than behavioural coding, as it has the advantage of easy and cost-effective implementation to achieve sufficient sample sizes/power and to draw inferences about populations. Currently, large-scale surveys have a strong focus on the frequency dimension, and teaching practices are mostly assessed through students' perceptions and experiences. For example, PISA asks students, the recipients of teaching practices, to report on the frequency of teachers' practices in classroom settings. Although such reporting does not tap into the quality or effectiveness of teaching practices, it provides data to test the structure and co-occurrence of teaching practices. However, as subjective teacher- or student-reports can be vulnerable to measurement bias, data quality and comparability across countries have to be tested (e.g., Vieluf, Kunter, & van de Vijver, 2013).

1.4.1. Multigroup confirmatory factor analysis (MGCFA)

Various psychometric tools are available to uncover the structure and metrics of self-reported teaching practices across countries. A conventional, rigorous approach involves multigroup confirmatory factor analysis (MGCFA). The common assumption with factor analysis is that items are indicators of latent factors and responses on items are "caused" by the latent factors. MGCFA provides a theory-driven approach (with a known factor structure) to examine a series of nested models across countries. These models include configural (i.e., the same configuration of zero and nonzero loadings of items on latent factors across countries), metric (i.e., the same factor loadings across countries), and scalar invariance (i.e., the same factor loadings and item intercepts across countries). Implications are attached with each level of invariance reached: configural invariance serves as a basis for any cross-country comparison, metric invariance allows valid comparisons of the unstandardized associations of constructs across countries (e.g., correlations between teaching practices and student outcomes), and only with scalar invariance can scale scores be compared across countries (i.e., means of teaching practices can be compared across countries, see Van de Vijver & Leung, 1997).

1.4.2. Network analysis

TALIS 2018 demonstrated that the different kinds of teaching practices are related (OECD, 2019) and, thus, it is important to consider the interdependency of teaching practices, without reducing them to a single "teaching practice score". Network analysis offers a novel perspective to gain insight into the co-occurrence (direct interactions) of observed indicators (Epskamp, Rhemtulla, & Borsboom, 2017) and helps us to understand how teaching practices are loosely or firmly related to each other as a system. It has been applied and is increasingly popular in personality research (see e.g., Costantini et al., 2015), research on political attitudes (see e.g., Dalege, Borsboom, van Harreveld, & Maas, 2018), and educational research (Abacioglu, Isvoranu, Verkuyten, Thijs, & Epskamp, 2019; Sachisthal et al., 2019). In contrast to factor analytic models, network analysis shifts the focus from the common shared variance to the variance between indicators (e.g., individual practices). Instead of assuming a common latent factor (e.g., "extraversion"), indicators (e.g., "I like to party", "I have a lot of friends") in a network are considered to mutually, directly affect each other – a change of one indicator leads to changes in the other connected indicator (e.g., by going to more parties, people meet more potential friends. And having more friends leads to more invitations to parties, see Costantini et al., 2015). Thus, indicators are part of the construct, instead of being measures of it (Sachisthal et al., 2019). The set of indicators (= *nodes*) and their interactions (= *edges*, representing unknown statistical relationships between two nodes) are visualized as a network, and magnitude (strength) and direction (positive versus negative) of pairwise interactions of indicators can be interpreted. Thus, network analysis illustrates if two teaching practices tend to co-occur frequently (positive relation) or not (absence of relations or negative relation) as well as the strength of their relation (if they are firmly or loosely related).

The function of indicators (i.e., nodes) within a network can be studied by examining their importance within the network (strength-centrality) or the structure and connectivity of the network. Nodes with higher values of *strength-centrality* influence other nodes more strongly than less central nodes, and thus, are the optimal starting point for targeted interventions or processes (Costantini et al., 2015). For teaching-practices, we view practices with the highest strength-centrality as the binding teaching practices (easily assigned with other practices) in the network. The *overall network structure* indicates the patterning of unique interactions between indicators in the network, and the *global connectivity* indicates the extent to which these indicators are connected (i.e., the extent to which teaching practices frequently co-occur) (Christensen, Kenett, Aste, Silvia, & Kwapil, 2018; Epskamp & Fried, 2018). The structure of a set of indicators and their overall connectivity can be compared across networks for different groups (i.e., different countries) by performing a network comparison test (NCT; Van Borkulo, Epskamp, & Milner, 2016).

1.4.3. Combining MGCFA and network analysis

Network analysis can complement MGCFA in several ways. First, although no latent factor is assumed or pursued in network analysis, clusters of indicators linked by strong edges may be indicative of latent factors underlying these indicators, making network analysis a diagnostic tool to explore the dimensionality of constructs. Secondly, network analysis focuses on the intricate interactions, providing a differing and additional nuanced look at the dynamics among indicators as a

system. Thirdly, comparisons based on network analysis do not require scalar invariance to be achieved. MGCFA aims to test whether individual or country means on the latent constructs can be compared validly (scalar invariance) and when scalar invariance is not tenable, the validity of further analysis on country means is not warranted (e.g., Vieluf et al., 2013). Yet, with many different countries included in a study, the shared core of a construct becomes smaller, making it nearly impossible to achieve scalar invariance (analysis paradox, see Van de Vijver, 2018). With network analysis, meaningful relations among items can be compared without pursing scalar measurement invariance. It has to be noted that measurement bias due to translation errors or different interpretations of the item content across groups (i.e., different countries) can nevertheless threat the validity and comparability of analysis results that are based on item responses (including network analysis).

1.5. The current study

We have summarized ongoing developments in the international debate on teaching practices and pointed out the possibility of country-specific structures and metrics in teaching practices, which cast doubts on the generalisability of a fixed structure of teaching practices. Nonetheless, empirical evidence on the structure of teaching practices across countries, with a focus on co-occurrence of teacher-directed, student-centred, and an addition of classroom assessment practices, is lacking. Thus, we formulate the following hypothesis and research question to explore the structure of teaching practices across countries.

Teacher-directed and student-centred teaching practices are based on well-founded theoretical approaches to instruction, particularly in Western countries. Moreover, both approaches have been operationalized and assessed in educational large-scale studies in dozens of countries (e.g., PISA, see OECD, 2014 and TALIS, see OECD, 2013), and they have guided designs to instruction across educational systems (e.g., Chile, see Zurita & Nussbaum, 2004 and Turkey, see Isikoglu, Basturk, & Karaca, 2009). Consequently, we expect to empirically identify the theoretically derived two distinct factors (i.e., teacher-directed versus student-centred practices) across countries (configural and metric invariance) (*Hypothesis 1)*. However, given the considerable influence exerted by culture and pedagogical traditions on the design of instruction, the likelihood that individuals from different countries understand and respond to this set of indicators in exactly the same way is low. Furthermore, unintended differences between cultures (e.g., how respondents make use of the response scale in the frequency-based measures) may further endanger cross-cultural comparability of scale scores on teaching practices obtained in large-scale educational surveys. Thus, we expect cross-country differences in item intercepts of teacher-directed and student-centred practices (no scalar invariance, *Hypothesis 2*), challenging the full scalar comparability of teaching practices. Despite its critical relevance, comparative research mostly compared teaching practices profiles across vastly different countries, without first demonstrating cross-cultural data comparability (e.g., OECD, 2013). This is an important omission that our study aims to remedy.

Unlike teacher-directed and student-centred practices, classroom assessment has been highlighted more recently in the international debate on teaching practices. Echazarra et al. (2016) positioned classroom assessment practices between traditional (teacher-directed) and modern (student-centred) approaches to instruction. Yet, it can be expected that specific assessment practices are infused in both types of practices to varying extents: Some assessment practices may show stronger relations to teacher-directed practices, whereas others might be more closely related to student-centred practices, and vice versa. For instance, Klieme (2020) suggests that formative assessment (including feedback) is more strongly related to teacher-directed instruction than to student-centred teaching. Similarly, effectiveness research often describes a combination of classroom assessment with either teacher-directed or student-centred practices (i.e., to structure or individualize instruction) as the most effective tool to boost student learning (OECD, 2013). Thus, operationalizing classroom assessment as a third, separate dimension of teaching practices (as for instance practiced by PISA, see OECD, 2014) might not be adequate. Instead, classroom assessment practices are expected to be compatible with both teacher-directed and student-centred ways of teaching. Yet, to date, most studies treated classroom assessment as latent-factor based (e.g., Klieme, 2020). Consequently, when integrating classroom assessment practices into teacher-directed and student-centred teaching activities, an alternative nuanced measurement model may be needed to unfold the intricate interactions between practices. With a wealth of data collected in large-scale educational assessment, such an attempt has not been made so far. The lack of theoretical foundation and empirical research thereby calls for an explorative approach. In this study, we explore how *individual* classroom assessment practices relate to teacher-directed and student-centred teaching practices across countries?

2. Method

2.1. Database and sample

We based our analysis on the 2012 cycle of PISA main study data of students' perceptions of teaching practices in mathematics lessons (see OECD, 2014). To ensure sufficient cultural variations and robustness in findings of different psychometric methods on the structure and metrics of teaching practices, we selected four clusters of countries based on main language families and included three countries/economies in each cluster. The selected country clusters not only differ in language, but also in their affluence level, cultural values of individualism-collectivism, and power distance, which have a bearing on the perceptions and preferences of teaching practices. Our chosen German- and English-speaking countries represent high affluence, high individualism, and low power distance cultures stemming from different pedagogical traditions (the German-speaking countries have highly tracked systems, the English-speaking countries have comprehensive school systems), while the Chinese- and Spanish-speaking countries represent moderately affluent, collectivistic, and high power distance cultures (the Spanish-speaking countries are infrequently studied in international comparisons to date and they add insight beyond the typical West-East comparisons) (Hofstede, 2001). To rule out method artefacts due to missing values and different sample sizes, a random subsample of 1000 students with complete responses on the targeted teaching practice items per country/economy were drawn. Therefore, analysis was conducted with 1000 students for each of three Chinese-speaking (Macao[2], Shanghai, Taipei), English-speaking (Australia, United Kingdom, United States), German-speaking (Austria, Germany, Switzerland), and Spanish-speaking (Chile, Colombia, Mexico) countries/economies, respectively (resulting in N = 12,000).

2.2. Measures

In the 2012 PISA, teaching practices encountered by students in mathematics lessons were measured with 13 items (five items for teacher-directed practices and four items each for student-centred and classroom assessment practices). Students responded on a 4-point Likert-type scale ranging from "*Every lesson*" to "*Never or hardly ever*" (see Table 1).

[2] 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.

Table 1
Items Measuring Teaching Practices in Mathematics Instruction (PISA 2012).

Practices	Item wording	Response scale
Teacher-directed	The teacher sets clear goals for our learning (T1).	1 = Every lesson
	The teacher asks me or my classmates to present our thinking or reasoning at some length (T2).	2 = Most lessons
	The teacher asks questions to check whether we have understood what was taught (T3).	3 = Some lessons
	At the beginning of a lesson, the teacher presents a short summary of the previous lesson (T4).	4 = Never or hardly ever
	The teacher tells us what we have to learn (T5).	
Student-Centred	The teacher gives different work to classmates who have difficulties learning and/or to those who can advance faster (S1).	
	The teacher assigns projects that require at least one week to complete (S2).	
	The teacher has us work in small groups to come up with joint solutions to a problem or task (S3).	
	The teacher asks us to help plan classroom activities or topics (S4).	
Classroom Assessment	The teacher tells me about how well I am doing in my mathematics class (A1).	
	The teacher gives me feedback on my strengths and weaknesses in mathematics (A2).	
	The teacher tells us what is expected of us when we get a test, quiz or assignment (A3).	
	The teacher tells me what I need to do to become better in mathematics (A4).	

2.3. Analysis strategy

These teaching practice items were analysed using both factor analysis (to test Hypotheses 1 and 2 with regard to teacher-directed and student-centred practices) and network analysis (to explore the relations of individual classroom assessment practices with teacher-directed and student-centred practices as formulated in the additional research question). All data and analysis codes are provided in the Open Science Framework .

2.3.1. Hypothesis-testing: identifying teacher-directed and student-centred teaching practices across countries

To test if teacher-directed and student-centred practices are two distinct factors across countries that reach metric invariance as postulated by Hypothesis 1 and 2, we first tested measurement invariance of a two-factor model comprising teacher-directed practices (five items) and student-centred practices (four items) in MGCFA across all 12 countries. Afterwards, we ran a three-factor MGCFA across the 12 countries to entertain the possibility that classroom assessment (measured with four items), next to the two factors in the first model, forms a third factor in the teaching practice framework. The model fit is evaluated by Chi-square tests and the Comparative Fit Index (CFI) (above 0.90 acceptable and above 0.95 excellent), the Root Mean Square Error of Approximation (RMSEA) (below 0.08 acceptable) and the Standardized Root Mean Square Residua (SRMR) (below 0.08 acceptable) (Cheung & Rensvold, 2002; Hu & Bentler, 1999). The acceptance of the more restricted model is based on the changes of CFI and RMSEA values in comparison to the less restricted model. In comparisons involving more than 10 groups, Rutkowski and Svetina (2014) proposed to set the cut point of change of CFI to 0.02 and that of RMSEA to 0.03 from the configural to the metric model, and from the metric to the scalar model the changes of both CFI and RMSEA should be within 0.01. We adhere to these criteria. All factor analyses were performed with the "lavaan" package in R (Rosseel, 2011).

2.3.2. Explorative approach: integrating classroom assessment into the framework of teaching practices

Next, we performed network analysis using the R-package *qgraph* (Epskamp, Cramer, Waldorp, Schmittmann, & Borsboom, 2012) to explore the structure and co-occurrence of teaching practices across countries. For each of the 12 countries, we estimated and visualized a partial correlation network (i.e., edges are estimated based on partial correlations between two indicators, controlling for all other indicators in the network, *cor_auto* was applied to create the correlation matrices). This analysis also incorporated a regression-based filtering approach, the "least shrinkage and selection operator" (LASSO), which leads to the estimation of a sparse, more interpretable model (with the hyperparameter gamma set to 0.50 for all models). Consequently, the absence of a connection (i.e., edge) represents conditional independence between two indicators (Christensen et al., 2018). To ensure the accuracy and stability of the estimates, a nonparametric bootstrapping test was performed for each country (i.e., for edges and the centrality index) (Epskamp & Fried, 2017). For these country-specific networks, we conducted three sets of analysis.

First, we performed pair-wise comparisons (= 66 comparisons) of the invariance of the *overall network structure* (operationalized as connection strength matrix) and the *global connectivity* (operationalized as weighted sum of absolute connections). They together inform about the similarity and differences of teaching practices with regard to global patterning across the country-specific networks. This was done with the significant testing based on permutations in the *NCT* package in R (Network Comparison Test, NCT, see Van Borkulo et al., 2016) with the LASSO regularization in which the hyperparameter was set to 0.50.

In each network, we compared the *edge differences* of the individual classroom assessment nodes to teacher-directed and student-centred practices, to clarify if individual classroom assessment practices are significantly more or less associated with (one of) the two established teaching practices. This was done in the R-package *bootnet* (Epskamp & Fried, 2017), with bootstrapped edge differences plotted out and their significance summarized.

Thirdly, we checked the similarity and differences in the importance of specific nodes in the country-specific networks. We focused on *strength-centrality* (sum of all edge weights connected to a given node in weighted networks). Nodes with higher values of strength-centrality influence other nodes more strongly, without considering other mediating nodes (Costantini et al., 2015). Other centrality indexes such as closeness and betweenness were not targeted, given their lower reliability and reproducibility in comparison to strength-centrality (Fried et al., 2018).

3. Results

3.1. Hypothesis-testing: identifying teacher-directed and student-centred teaching practices across countries

A MGCFA was performed on the nine items distinguishing teacher-directed and student-centred practices across all 12 countries. The model fit (see Table 2) points to acceptable metric invariance

Table 2
Model Fit of Measurement Invariance Tests for Teacher-Directed and Student-Centred Practices in Multigroup Confirmatory Factor Analysis.

	χ^2	df	CFI	RMSEA	SRMR
Configural	1882.911**	312	.927	.071	.045
Metric	*2178.765***	*389*	*.907*	.069	.070
Scalar	11635.861**	964	.740	.109	.089

Note. Most restrictive model with acceptable fit is printed in italics. **$p < .01$.

(acceptable CFI, RMSEA, and SRMR values for the metric model and drop of CFI and RMSEA values within the cut-off values of 0.02 and 0.03 respectively), which indicated a universal factor structure of teachers' practices with one factor for teacher-directed and one for student-centred practices, respectively. Hypothesis 1 was supported. In the metric invariance model, the factor loadings for teacher-directed practice items ranged from 0.54 to 0.63, and for the student-centred practice items from 0.55 to 0.64, suggesting that these items were relatively comparable indicators for the two constructs. However, scalar invariance was not supported, which was not unexpected (Hypothesis 2 supported). This may be due to intrinsic differences in metrics of these constructs across cultures or methodological artefacts that prevented valid cross-country comparisons on mean levels of the two constructs.

In the three-factor MGCFA model, in which teacher-directed and student-centred practices and classroom assessment were distinguished, only the configural model was just accepted [χ^2 (744, N = 12,000) = 5084.600, $p < 0.01$, CFI = 0.900, RMSEA = 0.076, SRMR = 0.050], and the factor loadings, item intercepts, and the associations among the three factors differed enormously across countries, pointing to a lack of support for the comparable three-factor solution across countries.

3.2. Explorative approach: integrating classroom assessment into the framework of teaching practices

To explore how classroom assessment practices relate to teacher-directed and student-centred practices across countries, partial correlation networks were estimated for each of the 12 countries (see Fig. 1a-l). Since the MGCFA for teacher-directed and student-centred practices demonstrated metric invariance, the construct scores (operationalized as the rounded mean scores across items measuring each construct) were used as nodes in the networks (nodes TD and SC) together with the classroom assessment items (Nodes A1-A4). These two construct scores can be considered ordered categories with the same metric as the classroom assessment items. The nonparametric bootstrapping testing based on 1000 bootstrapped samples showed support for the accuracy of the networks. The strength centrality indexes also showed acceptable stability, with the stability coefficients (CS cor = 0.70) all over 0.50 except for the US and Chile (both still over 0.25, with a value of 0.44 and 0.36, respectively). Supplement 1 presents all graphs for the recovery of the edges per country and Supplement 2 presents a table of the correlation stability coefficient [CS(cor = 0.70)] per country.

3.3. Overall network structure and global connectivity

A visual inspection of the 12 country-specific networks revealed that most edges were positive, indicating that the more frequent application of one practice seems to go hand in hand with the more frequent application of another connected practice, conditioning on all remaining practices. Even the TD and SC nodes were positively connected across countries, with relatively stronger edge weights in the networks for the Spanish- and German-speaking countries (weights between 0.19 and 0.27), and comparably weaker edge weights in the networks for the Chinese-speaking countries (weight Macao = 0.15 and Taipei = 0.17), and particularly in Shanghai (weight = 0.09). A few exceptions of negative edges (dashed lines) were observed in the networks for Taipei (SC and A4), the US and Austria (both SC and A3), and Chile and Mexico (both TD and A2), and all these edges were weak (weights between -0.06 and -0.10).

Table 3 presents the results of the pairs-wise tests of the overall network structure invariance (M-test statistics above the diagonal) and global connectivity invariance (S-test statistics below the diagonal). The overall network structure significantly differed for 48 of the 66 pair-wise comparisons ($p < 0.05$). Among countries belonging to the same linguistic/cultural cluster, the overall network structure invariance (i.e., comparability) was supported for all three English-speaking countries, Switzerland and Austria, Shanghai and Taipei, and Mexico and Chile. For countries belonging to different linguistic/cultural clusters, the network structures mostly differed. Exceptions were for Switzerland and all three English-speaking countries; Austria and the US and UK; and Macao and Chile and two German-speaking countries each (Macao: Austria and Germany; Chile: Austria and Switzerland). The network structure for Germany, Shanghai, Mexico, and Colombia showed the least comparability, with only one invariant comparison each. Even though not supported for all within-cluster pair-wise comparisons, it seems that there was a more similar network structure for countries belonging to the same linguistic/cultural cluster (especially the English-speaking cluster) than of countries belonging to different linguistic/cultural clusters.

With regard to the pair-wise comparisons of the global connectivity, the US network was significantly more connected (i.e., these teaching practices frequently co-occurred) than the networks for most other countries. The same applied to the network for Chile (compared to Austria, Shanghai, and Macao), and Mexico (compared to Germany, Shanghai, and Macao). The US network showed the comparably highest global strength (S = 2.55), followed by Mexico (S = 2.39), and Chile (S = 2.37), whereas the networks for the Chinese- and German-speaking countries showed a comparably low global connectivity (for global strength indices per country see Table 4). To summarize, these country-specific networks not only mostly significantly differed with regard to network structure (to a lesser extent for countries belonging to the same linguistic/cultural cluster), but also in global connectivity.

3.4. Relation of classroom assessment practices to TD and SC

We paid special attention to the individual classroom assessment nodes and their relations (i.e., edges) to the TD and SC nodes per country. A visual inspection suggested that the classroom assessment nodes did not cluster strongly, but showed rather distinct partial correlations with either TD or SC. For each country-specific network, bootstrapping was performed to test if individual classroom assessment practices significantly differently related to either TD or SC. The bootstrapped differences between all edge weights were plotted out and are presented in Supplement 3. We summarized the significance of edge differences between each of the four classroom assessment nodes and TD and SC in Table 5.

In all country-specific networks (except for Macao) A4 (informing individual students about what is needed to become better in mathematics) was more strongly, conditionally related to TD than SC. Similarly, A3 (informing what is expected of the class in tests or assignments) exhibited a significantly stronger unique relation with TD compared to SC in all countries, except in Shanghai and Mexico. The remaining two classroom assessment practices (A1: informing about the performance in mathematics class; A2: giving individual feedback on strength and weaknesses) showed some ambiguous relations to TD and SC. In all Chinese-speaking countries (as well as in Columbia), A1 was significantly more strongly, conditionally related to SC than TD, whereas in all other countries, these two edge differences were not significant. Among the non-Chinese speaking countries, A2 was more strongly conditionally linked with SC than TD (two other exceptions were Germany and Columbia, where no significant difference between the edge weights was observed). Thus, the four classroom assessment nodes did not cluster together strongly; they rather exhibited different relations to either teacher-directed or student-centred teaching practices, as detailed above.

3.5. Strength-centrality of individual nodes

In a next step, we investigated the strength-centrality of individual nodes within each country-specific network (see Fig. 2). Across countries, informing on individual strength and weaknesses in mathematics (A2) seemed to play a central role (average strength: 1.03) followed by the teacher-directed node (average strength: 0.96), and the two

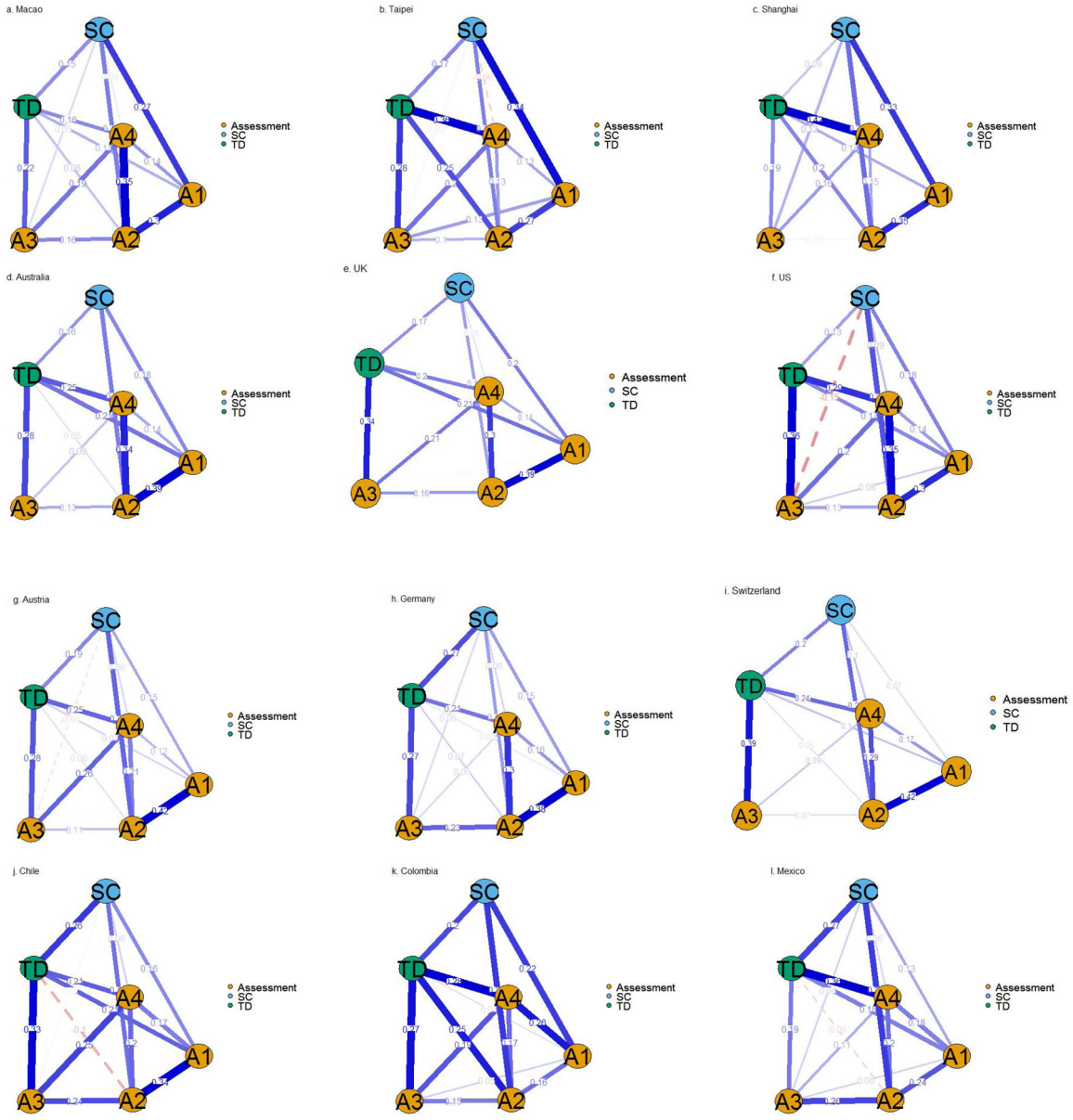

Fig. 1. Country-specific Partial-Correlation Networks of Teaching Practices.
Note. Partial correlation networks of teaching practices with rounded mean-scores for teacher-directed (TD) and student-centred (SC) practices, individual items for classroom assessment. Full (blue) lines represent positive edges; dashed (red lines) represent negative edges. To facilitate visualization, the position of the nodes is the same across networks (Germany is reference country). A1 = feedback performance in class, A2 = feedback individual strength and weaknesses, A3 = informing about expectations in test, A4 = feedback how to improve. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

assessment practices telling individual students what is needed to become better in mathematics (A4, average strength = 0.88), and informing students about how they are performing in their mathematics class (A1, average strength = 0.82). The student-centred and A3 node (what the class needs for a test, quiz, or assignment) played a less central role across countries (average strength: 0.67 and 0.69, respectively). The remaining nodes varied with regard to their importance across countries (particularly A3 with strength centrality values between 0.50 and 0.92). The country-specific strength-centrality of individual nodes can be found in Supplement 4.

4. Discussion

We set out to investigate the cross-cultural similarities and differences in the structure and co-occurence of teaching practices in mathematics instruction with a 12-country dataset from a large-scale international survey (PISA). We combined factor analysis and network analysis to test our hypothesis and research question. Rooted in instructionist and constructivist theories of teaching (Tobias & Duffy, 2010), the distinction between teacher-directed and student-centred teaching practices and their similar structure but not origin of metrics (i.e., item intercepts) across cultures were postulated (Hypothesis 1 and 2). Given the lack of theory and empirical foundation, we additionally explored how classroom assessment practices position within the broad range of teaching practices and investigated how individual assessment practices differently related to either teacher-directed or student-centred teaching practices.

We confirmed metric but not scalar invariance of teacher-directed and student-centred practices in the MGCFA of students' self-reported frequency of practices across countries (supporting Hypothesis 1 and 2).

Table 3
Results Pair-wise Network Comparison Test (NCT) for Network Structure and Global Connectivity Invariance.

	AUT	GER	CHE	AUS	UK	US	CNQ	TAP	MAC	COL	CHL	MEX
AUT		0.15**	0.12	0.13**	0.11	0.11	0.20**	0.15**	0.12	0.21**	0.10	0.17**
GER	0.06		0.15**	0.12**	0.13**	0.15**	0.19**	0.15**	0.10	0.17**	0.15**	0.14**
CHE	0.10	0.03		0.11	0.09	0.08	0.23**	0.19**	0.14**	0.19**	0.13	0.16**
AUS	0.09	0.03	0.01		0.10	0.09	0.16**	0.17**	0.10	0.18**	0.14**	0.15**
UK	0.11**	0.05	0.02	0.02		0.09	0.18**	0.19**	0.10**	0.19**	0.10	0.15**
US	0.34**	0.28**	0.25**	0.25**	0.23**		0.19**	0.19**	0.14**	0.19**	0.15**	0.17**
CNQ	0.07	0.01	0.03	0.02	0.04	0.27**		0.10	0.21**	0.20**	0.20**	0.20**
TAP	0.12	0.06	0.02	0.03	0.01	0.22**	0.05		0.16**	0.13	0.22**	0.22**
MAC	0.05	0.02	0.05	0.04	0.07	0.30**	0.02	0.07		0.15**	0.14**	0.15**
COL	0.11	0.05	0.02	0.02	0.00	0.23	0.04	0.01	0.07		0.22**	0.22**
CHL	0.16**	0.09**	0.06	0.07	0.04	0.19**	0.09	0.03	0.11**	0.04		0.12
MEX	0.18**	0.12**	0.08	0.09	0.07	0.16**	0.11**	0.06	0.13**	0.07	0.02	

Note. Above diagonal: network structure invariance (M-statistic), below diagonal: network global connectivity invariance (S-statistic); **= significantly different network structure, global connectivity if $p < 0.05$; AUS = Australia, AUT = Austria, CHE = Switzerland, CHL = Chile, COL = Colombia, GER = Germany, UK = Great Britain, MAC = Macao, MEX = Mexico, QCN = Shanghai, TAP = Taipei, US = United States.

Table 4
Global Network Connectivity per Country.

Country	Global strength
Macao	2.25
Taipei	2.33
Shanghai	2.28
Australia	2.30
UK	2.32
US	2.55
Austria	2.21
Germany	2.27
Switzerland	2.31
Chile	2.37
Colombia	2.32
Mexico	2.39

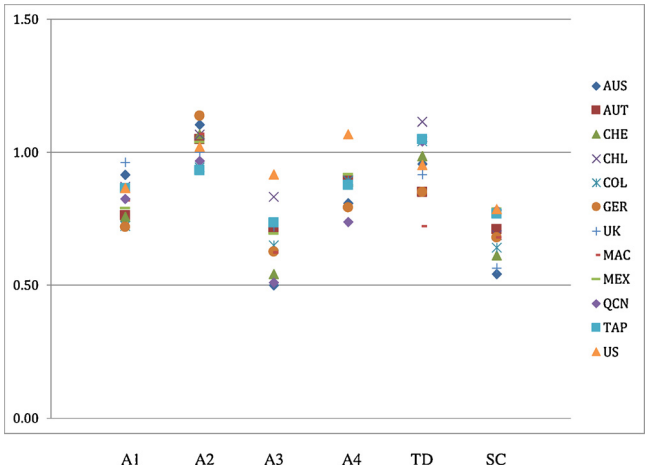

Fig. 2. Strength Index of the Partial-Correlation Networks across Countries.
Note. AUS = Australia, AUT = Austria, CHE = Switzerland, CHL = Chile, COL = Colombia, GER = Germany, UK = Great Britain, MAC = Macao, MEX = Mexico, QCN = Shanghai, TAP = Taipei, US = United States. A1= feedback performance in class, A2= feedback individual strength and weaknesses, A3= informing about expectations in test, A4= feedback how to improve, TD = rounded mean score for teacher-directed practices, SC = rounded mean score for student-centred practices.

Adding classroom assessment as a third factor in the MGCFA did not support an invariant three-factor structure across countries; whereas a network analysis per country on individual classroom assessment practices and the rounded mean scores of teacher-directed and student-centred practices showed rather different direct interactions among the teaching practices. Network analyses revealed that (1) across countries, most teaching practices were positively mutually linked (even teacher-directed and student-centred practices), (2) the overall network structure and to a lesser extent global connectivity differed for most pair-wise comparisons, but similarity of the network structure was often found for countries belonging to the same cultural and linguistic cluster; (3) the classroom assessment items did not form a cluster and do not seem to be latent-factor based and among the four classroom assessment practices, A4 (informing individual students about what is needed to become better in mathematics) and A3 (informing what is expected of the class in tests or assignments) more strongly related to teacher-directed practices than student-centred practices across countries, whereas the other two classroom assessment practices (A1: informing on how the student is doing in the mathematics class; A2: feedback on individual strength and weaknesses) showed less common patterning in their relation to either teacher-directed or student-centred practices, but tended to be more mutually linked to student-centred practices, and (4) in comparisons of the relative importance of specific practices in the country-specific networks, A2 (informing on individual strength and weaknesses in mathematics) and the node for teacher-directed practices played a relative important role on average across countries, whereas the node for student-centred practices was less important. In the following, we discuss the global patterning and

Table 5
Significance of Edge Difference Tests of Each Classroom Assessment Node (A) with Teacher-Directed (TD) and Student-Centred (SC) Nodes.

Edges compared	AUT	GER	CHE	AUS	UK	US	CNQ	TAP	MAC	COL	CHL	MEX
A1-TD vs A1-SC	X	X	X	X	X	X	V	V	V	V	X	X
A2-TD vs A2-SC	V	X	V	V	V	V	X	X	X	X	V	V
A3-TD vs A3-SC	V	V	V	V	V	V	X	V	V	V	V	X
A4-TD vs A4-SC	V	V	V	V	V	V	V	V	X	V	V	V

Note. V indicates significant edge difference at $p < 0.05$; X indicates nonsignificant edge difference at $p < 0.05$. A1= feedback performance in class, A2= feedback individual strength and weaknesses, A3= informing about expectations in test, A4= feedback how to improve, TD = rounded mean score for teacher-directed practices, SC = rounded mean score for student-centred practices. AUS = Australia, AUT = Austria, CHE = Switzerland, CHL = Chile, COL = Colombia, GER = Germany, UK = Great Britain, MAC = Macao, MEX = Mexico, QCN = Shanghai, TAP = Taipei, US = United States.

implications. We refrain from diving into specifics of cross-country differences, given that no clear expectation was formulated and the exploratory nature of the analysis.

4.1. Teacher-directed and student-centred practices: two distinct approaches to teaching?

Theoretically, teacher-directed and student-centred practices are based on two distinct - and even often labelled as opposite - approaches to instruction. Our MGCFA supports this theoretical distinction across countries. Thus, the theories of instruction developed and tested in Western countries are generalizable to the non-Western countries in our study (e.g., East-Asian and Latin-American countries). However, the consistently positive conditional relation between teacher-directed and student-centred practices in our country-specific networks highlights that teachers do not stick to only one approach to teaching, but combine practices stemming from different teaching traditions. Even within a lesson students are likely to be exposed to various teaching practices (Echazarra et al., 2016). Thus, teacher-directed and student-centred practices complement each other to fit the context, subject content, and students. Consequently, the strict theoretical distinction might not reflect the more flexible co-occurrence of teaching practices in reality. This seems to be less the case for the Chinese-speaking countries, where we observed a comparably low, yet positive relation between teacher-directed and student-centred practices (i.e., less frequent co-occurrence of teacher-directed and student-centred practices). One possible explanation is that East-Asian countries value conformity and legitimize power distance more than the other linguistic/cultural clusters of countries (Hofstede, 2001), thus they tend to strictly adhere to one specific instructional approach (i.e., traditional teacher-directed instruction, see Echazarra et al., 2016).

4.2. Integrating classroom assessment into the framework of teaching practices

Our network analysis on the structure and co-occurrence of teaching practices challenges the proposal to conceptualize classroom assessment practices as a third set of practices as well as the positioning between traditional (teacher-directed) and modern (student-centred) approaches to teaching (Echazarra et al., 2016). This characterisation might be an oversimplification of the complex nature of classroom assessment. Instead of clustering together, these individual classroom assessment practices tended to show a stronger relation to either teacher-directed or student-centred practices. A more teacher-directed approach to instruction is clearly related to assessment practices that are used to structure and guide classroom learning, such as informing students about learning goals (i.e., what is expected in tests, a quiz, or assignments) or providing advice on how to reach specific goals (i.e., what is needed to become better in mathematics). A more student-centred approach to instruction, on the other hand, tends to be related to assessment practices supporting individualized learning, such as providing individual feedback on strength and weaknesses or feedback with a social reference frame (i.e., how well a student is doing in mathematics class). Thus, network analysis provides a more nuanced look on the relation between individual assessment practices and teacher-directed or student centred-practices. It should be noted that both directions of the relation are possible: i.e., the specific approaches to teaching lead to the co-occurrence of specific assessment practices or vice versa. Consequently, it is plausible that these classroom assessment practices do not stem from one tradition, but are instilled in teaching from multiple traditions. Moreover, cross-cultural differences on strength of the links add complexity to the picture. In any case, treating them as one factor would obscure these nuanced differences. Moreover, our results emphasize the broad nature of the concept *teaching practices*, intertwining practices stemming from multiple teaching traditions with a complex relation to each other. We urge further research to define the theoretical concept more precisely.

4.3. The structure and dynamics of teaching practices across countries

Across countries, we found mostly different network structures and global connectivity. However, we also found an invariant structure of the networks among the three English-speaking countries; Taipei and Shanghai; Austria and Switzerland; and Mexico and Chile, indicating more similarity within linguistic/cultural clusters of countries than across clusters (this is in line with findings of Fischer, Praetorius, & Klieme, 2019). The cultural and colonial heritage of the three English speaking countries, their shared teaching traditions and (comprehensive) school system structure seem to be more similar than in the other linguistic/cultural clusters, which may contribute to higher levels of similarity of the networks (the comparability of teaching constructs for English-speaking countries was also demonstrated in other studies, see Fischer et al., 2019 or Klieme, 2020). Across linguistic/cultural clusters, interestingly, Switzerland's network structure was comparable to the structure of all three English-speaking countries, and also Austria's network showed an invariant structure compared to the UK and US. Thus, German- and English- speaking countries might be relatively similar in terms of teaching culture, compared to Chinese- and Spanish-speaking countries. To draw valid conclusions, future research should investigate similarities and differences between countries in more detail (e.g., with regard to school systems, preferred teaching approaches, but also cultural and colonial traditions). Moreover, countries also differ with regard to the importance of different teaching and assessment practices and the relation of individual assessment practices with either teacher-directed or student-centred practices in particular (see previous section). Our results emphasize context-specific structures and patterns of teaching practices. Consequently, targeted interventions have to be tailored to the specific context in order to be effective in the respective countries and should not be overly generalized or "borrowed" across countries.

4.4. Centrality of teaching practices: starting point for targeted interventions

With strength centrality indices, we also witness the relative importance of individual practices within a network of teaching practices. We view the most central practices as the binding practices in the teaching practice networks. In other words, they are versatile because they can accompany many other practices and are easily aligned with other practices. It is our extrapolation that increasing the central practice may facilitate promoting other practices to be used in combination. We thus expect that a stimulation of the most central practice is beneficial as it might influence many other practices that are well linked with it. Classroom assessment practices with a focus on individual students - particularly the practice of providing individualized feedback on strength and weaknesses (A2, most important node on average across countries) and the practice of giving individual advice on how to get better (A4) seem to be at the heart of teaching as perceived by students. In contrast, assessment practices focusing on the class (A3: the teacher tells us what is expected in a quiz or assignment, A1: the teacher compares me with my class) seem to be less influential on other teaching practices. Similarly to findings of Echazarra et al. (2016), teacher-directed practices seem to be more important in mathematics instruction than student-centred practices in our study.

4.5. Toolbox for measurement investigations

Methodologically, empirically testing measurement invariance of constructs in large-scale surveys before drawing any cross-cultural comparison is important in order to ensure the level of comparability and draw valid comparative inferences (Boer, Hanke, & He, 2018). Psychometric tools abound (e.g., item response theory-based scaling, latent class analysis), and flexible applications are in much need. We

made use of two methods for different purposes. MGCFA was used for theory testing and confirmation, whereas network analysis was resorted to aid measurement in exploratory ways. MGCFA with its various adaptations and extensions (e.g., partial invariance, approximate invariance testing, or simultaneous mixture CFA) provides rigorous and realistic testing of measurement of multiple-item measures. Network analysis is especially useful for relatively new constructs with less clear conceptualizations (e.g., classroom assessment practices) and that may not be latent-factor based (Costantini et al., 2015). These tools complement each other and deepen our understanding on substantive educational phenomena, as they either capture the commonality (MGCFA) or the unique interactions not accounted by the commonality (network analysis). For network analysis, there is a new development towards a better integration with classic psychometrics (Epskamp et al., 2017), and new research questions can be answered with information gathered in network analysis (e.g., what combination or dynamics of teaching practices especially contribute to student learning, how global connectivity in partial correlation networks of teaching practices is related to national policy on teacher autonomy).

4.6. Limitations

When interpreting the results of our study, some limitations have to be considered. Firstly, we used PISA data, where students are nested within schools (without clustering at classroom level). Self-reports of students taught in possibly different classrooms by different teachers ignore the heterogeneity at classroom levels, and thus have inferential limits. 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, Robitzsch, Trautwein, & Kunter, 2009). Future research should use multiple data sources (especially teachers' self-report and observations in real classes) to validate our results. Secondly, potential measurement bias in item responses (e.g., translation errors, misinterpretation of item content) may be detected in MGCFA, but still may exist in network analysis, which can challenge the validity of comparisons of structure, edge weights, and centrality indices across countries. Other psychometric tools and qualitative procedures are in need to further uncover bias that can limit data comparability. Thirdly, to facilitate comparisons, we randomly selected 1000 students per country. Replications with different subsamples per country or additional country clusters may be performed to check the robustness of our results. And lastly, following the results of the MGCFA (identifying two separate factors across countries) - we included teacher-directed and student-centred teaching practices as rounded mean construct scores in our network analysis. Thus, we had no information on which specific teacher-directed and student-centred practices are interlinked and the strength of their connection. Further research should investigate under which circumstances teachers combine which teaching practices as well as the effectiveness.

5. Conclusion

We have made use of data of representative student samples from multiple countries and complementary psychometric methods to study the structure and co-occurrence of teaching practices from a cross-cultural perspective. Our empirical support for the distinction between teacher-directed and student-centred practices, and the nuanced differences in classroom assessment practices related to these two well-established teaching practices open up for new perspectives to conceptualize dimensions of teaching practices. We urge researchers to apply innovative measurement models, and expand the measurement to other facets beyond the quantitative focus.

Funding

This work was partially supported by the Marie Sklodowska-Curie Individual Fellowship European program [grant number 748788].

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.stueduc.2020.100861.

All data and syntax used in this study are available at https://osf.io/e4fx6/.

References

Abacioglu, C. S., Isvoranu, A.-M., Verkuyten, M., Thijs, J., & Epskamp, S. (2019). Exploring multicultural classroom dynamics: A network analysis. *Journal of School Psychology, 74*. https://doi.org/10.1016/j.jsp.2019.02.003.

Astin, W. A., Banta, W. T., Cross, P. K., El-Khawas, E., Ewell, T. P., Hutchings, P., et al. (1996). *Nine principles of good practice for assessing student learning.* American Association for Higher Education (AAHE).

Boer, D., Hanke, K., & He, J. (2018). On detecting systematic measurement error in cross-cultural research: A review and critical reflection on equivalence and invariance tests. *Journal of Cross-Cultural Psychology, 49*, 713–734. https://doi.org/10.1177/0022022117749042.

Fischer, J., Praetorius, A.-K., & Klieme, E. (2019). The impact of linguistic similarity on cross-cultural comparability of students' perceptions of teaching quality. *Educational Assessment Evaluation and Accountability, 31*, 201–220. https://doi.org/10.1007/s11092-019-09295-7.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment Evaluation and Accountability, 21*, 5–31. https://doi.org/10.1007/s11092-008-9068-5.

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. https://doi.org/10.1016/j.stueduc.2016.03.005.

Carroll, J. B. (1963). A model of school learning. *Teachers College Record, 64*, 723–733.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233–255. https://doi.org/10.1207/s15328007sem0902_5.

Christensen, A. P., Kenett, Y. N., Aste, T., Silvia, P. J., & Kwapil, T. R. (2018). Network structure of the Wisconsin Schizotypy Scales-Short Forms: Examining psychometric network filtering approaches. *Behavior Research Methods, 50*, 2531–2550. https://doi.org/10.3758/s13428-018-1032-9.

Coombs, A., DeLuca, C., LaPointe-McEwan, D., & Chalas, A. (2018). Changing approaches to classroom assessment: An empirical study across teacher career stages. *Teaching and Teacher Education, 71*, 134–144. https://doi.org/10.1016/j.tate.2017.12.010.

Costantini, G., Epskamp, S., Borsboom, D., Perugini, M., Mõttus, R., Waldorp, L. J., et al. (2015). State of the aRt personality research: A tutorial on network analysis of personality data in R. *Journal of Research in Personality, 54*, 13–29. https://doi.org/10.1016/j.jrp.2014.07.003.

Creemers, B. P. M., & Kyriakides, L. (2006). Critical analysis of the current approaches to modelling educational effectiveness: The importance of establishing a dynamic model. *School Effectiveness and School Improvement, 17*, 347–366. https://doi.org/10.1080/09243450600697242.

Dalege, J., Borsboom, D., van Harreveld, F., & Maas, H. (2018). A network perspective on attitude strength: Testing the connectivity hypothesis. *Social Psychological and Personality Science, 10*, 746–756. https://doi.org/10.1177/1948550618781062.

Dewey, J. (1929). *My pedagogic creed.* Washington, DC: Progressive Education Association.

Duncker, K., & Lees, L. S. (1945). On problem-solving. *Psychological Monographs, 58*, i–113. https://doi.org/10.1037/h0093599.

Echazarra, A., Salinas, D., Méndez, I., Denis, V., & Rech, G. (2016). *How teachers teach and students learn: Successful strategies for school. OECD Education Working Paper.* Paris: OECD Publishing.

Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods, 23*, 617–634. https://doi.org/10.1037/met0000167.

Epskamp, S., & Fried, E. I. (2017). *Bootnet: Bootstrap methods for various network estimation routines.* Retrieved fromhttps://cran.r-project.org/web/packages/bootnet/index.html.

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized network psychometrics: Combining network and latent variable models. *Psychometrika, 82*, 904–927. https://doi.org/10.1007/s11336-017-9557-x.

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). Qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software, 48*, 1–18. https://doi.org/10.18637/jss.v048.i04.

Fried, E. I., Eidhof, M. B., Palic, S., Costantini, G., Huisman-van Dijk, H. M., Bockting, C. L. H., et al. (2018). Replicability and generalizability of Posttraumatic Stress Disorder (PTSD) networks: A cross-cultural multisite study of PTSD symptoms in four trauma patient samples. *Clinical Psychological Science, 6*, 335–351. https://doi.org/10.1177/2167702617745092.

Fuller, B., & Clarke, P. (1994). Raising school effects while ignoring culture? Local conditions and the influence of classroom tools, rules, and pedagogy. *Review of Educational Research, 64*, 119–157. https://doi.org/10.3102/00346543064001119.

Harlen, W. (2007). Formative classroom assessment in science and mathematics. In J. H. McMillan (Ed.). *Formative classroom assessment: Theory into practice* (pp. 116–135). New York/London: Teachers College Press, Columbia University.

Harlen, W., & Deakin-Crick, R. (2002). *A systematic review of the impact of summative assessment and tests on students' motivation for learning.* London: EPPI-Centre.
Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* London: Routledge.
Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations.* Thousand Oaks, CA: Sage.
Hu, L.t., & 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. https://doi.org/10.1080/10705519909540118.
Isikoglu, N., Basturk, R., & Karaca, F. (2009). Assessing in-service teachers' instructional beliefs about student-centered education: A Turkish perspective. *Teaching and Teacher Education, 25*, 350–356. https://doi.org/10.1016/j.tate.2008.08.004.
Jacobs, J. K., Hollingsworth, H., & Givvin, K. B. (2007). Video-based research made "easy": Methodological lessons learned from the TIMSS video studies. *Field Methods, 19*, 284–299. https://doi.org/10.1177/1525822X07302106.
Kippers, W. B., Wolterinck, C. H. D., Schildkamp, K., Poortman, C. L., & Visscher, A. J. (2018). Teachers' views on the use of assessment for learning and data-based decision making in classroom practice. *Teaching and Teacher Education, 75*, 199–213. https://doi.org/10.1016/j.tate.2018.06.015.
Klieme, E. (2020). Policies and practices of assessment: A showcase for the use (and misuse) of International Large Scale Assessments in Educational Effectiveness Research. In J. Hall, P. Sammons, & A. Lindorff (Eds.). *International Perspectives in Educational Effectiveness Research*. Springer.
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.
McCormick, J., & Alavi, S. B. (2004). A cross-cultural analysis of the effectiveness of the Learning Organization model in school contexts. *International Journal of Educational Management, 18*, 408–416. https://doi.org/10.1108/09513540410563112.
Mostafa, T., Echazarra, A., & Guillo, H. (2018). *The science of teaching science: An exploration of science teaching practices in PISA 2015. OECD Education Working Paper.* Paris: OECD Publishing.
OECD (2013). *Teaching and learning international survey TALIS 2013: Conceptual framework.* Paris: OECD Publishing.
OECD (2014). *PISA 2012 technical report*Paris: OECD Publishing.
OECD (2015). *PISA 2015 assessment and analytical framework.* Paris: OECD Publishing.
OECD (2019). *Teaching and learning international survey TALIS 2018: Technical report*Paris: OECD Publishing.
Ormrod, J. E. (2012). *Essentials of educational psychology: Big ideas to guide effective teaching* (3rd ed.). Boston: Pearson.
Piaget, J. (1952). *The origins of intelligence in children.* New York, NY: International Universities Press.
Praetorius, A.-K., Klieme, E., Bell, C. A., Qi, Y., Witherspoon, W., & Opfer, D. (2018). *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*.
Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science, 28*, 4–13. https://doi.org/10.1002/bs.3830280103.
Richardson, V. (2003). Constructivist pedagogy. *Teachers College Record, 105*, 1623–1640. https://doi.org/10.1046/j.1467-9620.2003.00303.x.
Rosenshine, B. (1976). Classroom instruction. In N. L. Gage (Ed.). *The psychology of teaching methods* (pp. 335–371). (75th ed.). Chicago, IL: University of Chicago Press.
Rosseel, Y. (2011). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*. https://doi.org/10.18637/jss.v048.i02.
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. https://doi.org/10.1177/0013164413498257.
Sachisthal, M. S. M., Jansen, B. R. J., Peetsma, T. T. D., Dalege, J., van der Maas, H. L. J., & Raijmakers, M. E. J. (2019). Introducing a science interest network model to reveal country differences. *Journal of Educational Psychology, 111*, 1063–1080. https://doi.org/10.1037/edu0000327.
Stigler, J., & Hiebert, J. (1999). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom.* New York, NY: The Free Press.
Tobias, S., & Duffy, T. M. (Eds.). (2010). *Constructivist instruction: Success or failure?*New York, London: Routledge Taylor & Francis Group.
Van Borkulo, C. D., Epskamp, S., & Milner, A. (2016). *NetworkComparisonTest.* https://cran.rproject.org/web/packages/NetworkComparisonTest/NetworkComparisonTest.pdf.
Van de Vijver, F. J. R. (2018). *Talk at the OECD-GESIS seminar: Translating and adapting instruments in large-scale assessments.* Paris.
Van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research.* Thousand Oaks: Sage.
Vieluf, S., Kaplan, D., Klieme, E., & Bayer, S. (2012). *Teaching practices and pedagogical innovations. Evidence from TALIS.* Paris: OECD Publishing.
Vieluf, S., Kunter, M., & van de Vijver, F. J. R. (2013). Teacher self-efficacy in cross-national perspective. *Teaching and Teacher Education, 35*, 92–103. https://doi.org/10.1016/j.tate.2013.05.006.
Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.
Zurita, G., & Nussbaum, M. (2004). A constructivist mobile learning environment supported by a wireless handheld network. *Journal of Computer Assisted Learning, 20*, 235–243. https://doi.org/10.1111/j.1365-2729.2004.00089.x.

Appendix D: Lebenslauf

Jessica Fischer

Wissenschaftlicher Werdegang

Seit 09/2020	**Wissenschaftliche Mitarbeiterin** Deutsches Institut für Erwachsenenbildung (DIE) - Leibniz-Zentrum für Lebenslanges Lernen, Bonn, Abteilung: Lehren, Lernen, Beraten, Projekt: Teachers in Adult Education - A Panel Study (TAEPS)
05/2016-09/2020	**Doktorandin** DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation, Frankfurt am Main, Abteilung: Bildungsqualität und Evaluation, Projekt: TALIS-Videostudie international, Doktorvater: Prof. Dr. Eckhard Klieme
05.2019	**Forschungsaufenthalt** Research Institute for International and Comparative Education, Shanghai Normal University, China
10/2015 - 02/2016	**Tutorin** Otto-Friedrich-Universität Bamberg, Lehrstuhl für Soziologie, insbesondere Sozialstrukturanalyse, Vorlesung: Einführung in die Migrationssoziologie, Prof. Dr. Cornelia Kristen
02/2013 - 07/2014	**Studentische Hilfskraft** Otto-Friedrich-Universität Bamberg, Lehrstuhl Soziologie I, Projekt: Diskriminierung im Bildungswesen - Mikromechanismen und Makrodeterminanten
09 – 10/2012	**Praktikum** Leibniz-Institut für Bildungsverläufe e.V. Bamberg, Abteilung: Methoden, Gewichte und Imputation
09/2011 – 04/2016	**Studentische Hilfskraft** Leibniz-Institut für Bildungsverläufe e.V. Bamberg, Abteilung: Methoden, Gewichte und Imputation

Studium

10/2017- 03/2020	**Ergänzungsstudium Psychologie** Goethe-Universität Frankfurt
10/2013 – 04/201	**Master of Arts – Soziologie** Otto-Friedrich-Universität Bamberg
10/2010 - 09/2013	**Bachelor of Arts – Soziologie**

Otto-Friedrich-Universität Bamberg

Publikationen

Zeitschriftenbeiträge (peer-reviewed)

1. **Fischer, J.,** Klieme, E., Praetorius, A.-K., & Jinjie, X. (submitted). Understanding lack of equivalence in cross-cultural measurements of teaching quality: Students' interpretations of student support items in Germany and China. *Submitted to Teaching and Teacher Education.*

2. **Fischer, J.**, He, J., & Klieme, E. (2020). The structure of teaching practices across countries: A combination of factor analysis and network analysis. *Studies in Educational Evaluation*, 65. https://doi.org/10.1016/j.stueduc.2020.100861

3. He, J. & **Fischer, J.** (2020). Differential associations of school practices with achievement and sense of belonging of immigrant and non-immigrant students. *Journal of Applied Developmental Psychology, 66*. https://doi.org/10.1016/j.appdev.2019.101089

4. **Fischer, J.,** Praetorius, A.-K., & Klieme, E. (2019). The impact of linguistic similarity on cross-cultural comparability of students' perceptions of teaching quality. *Educational Assessment, Evaluation and Accountability, 31*. https://doi.org/10.1007/s11092-019-09295-7

Buchbeiträge (peer-reviewed)

1. He, J., Buchholz, J., & **Fischer, J**. (accepted). Cross-cultural comparability of latent constructs in ILSAs. In T. Nilsen, A. Stancel-Piątak, & J.-E. Gustafsson (Eds.), *Springer International Handbooks of Education. International Handbook of Comparative Large-Scale Studies in Education: Perspectives, Methods and Findings.* Springer.
2. He, J. & **Fischer, J.** (2018). Methods for measurement invariance testing for contextual scales in large-scale educational assessments (pp. 76-88). In C. Magno & A.P. David (Eds.), *Philippine and global perspectives on educational assessment.* Philippine Educational Measurement and Evaluation Association.

Vorträge und Poster auf wissenschaftlichen Tagungen und Kongressen

Vorträge (peer-reviewed)

1. **Fischer, J.,** He, J., & Klieme, E. (August, 2019). *The structure of teaching practices across countries: A combination of factor analysis and network analysis.* Vortrag im Rahmen der 18. Tagung der European Association for Research on Learning and Instruction (Earli) in Aachen.

2. **Fischer, J.,** He, J., & Klieme, E. (April, 2019). *The structure of teaching practices across countries: A combination of factor analysis and network analysis.* In J. He (Chair), Teaching practices from a cross-cultural perspective: Methodological rigors and innovations. Symposium im Rahmen der 63. Jahrestagung der Comparative & International Educational Society (CIES) in San Francisco, USA.

3. **Fischer, J.**, Praetorius, A.-K., & Klieme, E. (August, 2018). *Measuring instructional quality: Invariance across and within linguistic groups.* In **J. Fischer** (Chair), Data comparability as a prerequisite for the evaluation of learning and teaching across groups and time. Symposium im Rahmen der Earli SIG 18 "Educational Effectiveness" & SIG 23 "Educational Evaluation, Accountability and School Improvement" in Groningen, Niederlande.

4. **Fischer, J.,** Praetorius, A.-K., & Klieme, E. (September, 2018). *The impact of linguistic similarity on measurement invariance of items measuring instructional quality.* In J. Buchholz (Chair), Invariance in international large-scale assessments: Methodological issues and empirical applications. Symposium im Rahmen der 51. Jahrestagung der deutschen Gesellschaft für Psychologie (DGPs) in Frankfurt.

5. **Fischer, J.**, Praetorius, A.-K., & Klieme, E. (Juli, 2018). *Measuring instructional quality: Invariance across and within linguistic groups.* In J. He (Chair), Enhancing data comparability in cross-cultural assessments. Symposium im Rahmen der 24. Jahrestagung der International Association for Cross-Cultural Psychology (IACCP) in Guelph, Kanada.

6. **Fischer, J**., Praetorius, A.-K., & Klieme, E. (Februar, 2018). *Internationale Vergleichbarkeit von Items zur Erfassung von Unterrichtsqualität*. In C. Köhler & A.-K. Praetorius (Chair), Wirkungen von Unterrichtsqualität: Bisherige Forschungsbefunde und methodische Herausforderungen. Symposium im Rahmen der 6. Jahrestagung der Gesellschaft für empirische Bildungsforschung (GEBF) in Basel, Schweiz.

7. **Fischer, J.**, Praetorius, A.-K., & Klieme, E. (November, 2017). *International comparability of student ratings for the assessment of instructional quality.* Vortrag im Rahmen des GESIS Symposiums "Advances in scale development in the social sciences: Issues of comparability" in Mannheim.

8. **Fischer, J.**, Praetorius, A.-K., & Klieme, E. (November, 2017). *Internationale Vergleichbarkeit von Schülerratings zur Erfassung von Unterrichtsqualität.* Vortrag bei der gemeinsamen Tagung der Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e.V. (ASI) und der Sektion Methoden der empirischen Sozialforschung der Deutschen Gesellschaft für Soziologie (DGS), Robert Koch-Institut in Berlin.

9. **Fischer, J.**, Praetorius, A.-K., & Klieme, E. (September, 2017). *Internationale Vergleichbarkeit von Schülerratings zur Erfassung von Unterrichtsqualität.* Vortrag bei der Pre-Conference der gemeinsamen Tagung der Fachgruppen Entwicklungspsychologie und Pädagogische Psychologie (PAEPSY) in Münster.

10. **Fischer, J.**, Praetorius, A.-K., & Klieme, E. (September, 2017). *Internationale Vergleichbarkeit von Schülerratings zur Erfassung von Unterrichtsqualität.* Vortrag bei der Pre-Conference der gemeinsamen Tagung der Arbeitsgruppe für Empirische Pädagogische Forschung (AEPF) und Kommission Bildungsplanung, Bildungsorganisation und Bildungsrecht (KBBB) in Tübingen.

Poster-Präsentationen (peer-reviewed)

1. **Fischer, J.,** Klieme, E., & Praetorius, A.-K. (Februar, 2019). *Erfassung von Unterrichtsqualität in China und Deutschland - Eine Mixed-Methods-Studie.* Poster-Präsentation bei der 7. Jahrestagung der Gesellschaft für empirische Bildungsforschung (GEBF) in Köln.

2. **Fischer, J.**, Praetorius, A.-K., & Klieme, E. (November, 2017). *Eine Kombination quantitativer und qualitativer Verfahren zur Beantwortung der Frage: Sind Schülerratings zur Erfassung von Unterrichtsqualität international vergleichbar?* Poster-Präsentation bei der IDeA Winter School "Bildungsforschung intermethodisch und interdisziplinär: Perspektiven für den wissenschaftlichen Nachwuchs" in Frankfurt.

Zeitschriften-Gutachtertätigkeiten

1. Educational Assessment, Evaluation and Accountability
2. Studies in Educational Evaluation

Eigenständige Leitung von Lehrveranstaltungen

09/2019	Cross-Cultural Research Methods (mit Dr. Jia He), zweitägiger Workshop in englischer Sprache im Rahmen der Methodenwoche der Goethe-Universität Frankfurt