Educational Assessment, Evaluation and Accountability (2020) 32:275–310
https://doi.org/10.1007/s11092-020-09328-6

Do student ratings provide reliable and valid information about teaching quality at the school level? Evaluating measures of science teaching in PISA 2015

Anindito Aditomo [1,2] · Carmen Köhler [2]

Received: 2 October 2019 / Accepted: 13 July 2020 / Published online: 17 July 2020

Abstract

Large-scale educational surveys, including PISA, often collect student ratings to assess teaching quality. Because of the sampling design in PISA, student ratings must be aggregated at the school level instead of the classroom level. To what extent does school-level aggregation of student ratings yield reliable and valid measures of teaching quality? We investigate this question for six scales measuring classroom management, emotional support, inquiry-based instruction, teacher-directed instruction, adaptive instruction, and feedback provided by PISA 2015. The sample consisted of 503,146 students from 17,678 schools in 69 countries/regions. Multilevel CFA and SEM were conducted for each scale in each country/region to evaluate school-level reliability (intraclass correlations 1 and 2), factorial validity, and predictive validity. In most countries/regions, school-level reliability was found to be adequate for the classroom management scale, but only low to moderate for the other scales. Examination of factorial and predictive validity indicated that the classroom management, emotional support, adaptive instruction, and teacher-directed instruction scales capture meaningful differences in teaching quality between schools. Meanwhile, the inquiry scale exhibited poor validity in almost all countries/regions. These findings suggest the possibility of using student ratings in PISA to investigate some aspects of school-level teaching quality in most countries/regions.

Keywords Teaching effect · Instructional quality · School climate · Multilevel modelling · Confirmatory factor analysis

✉ Anindito Aditomo
aditomo@dipf.de; aditomo@staff.ubaya.ac.id

[1] Faculty of Psychology, University of Surabaya, Surabaya, Indonesia

[2] DIPF | Leibniz Institute for Research and Information in Education, Rostocker Str. 6, 60323 Frankfurt am Main, Germany

1 Introduction

Large-scale assessments of learning have shaped public discourse and influenced educational policy in many countries (Breakspear 2012; Grek 2009). While attention has mostly focused on students' achievement, large-scale assessment studies also provide rich information regarding the school context and educational processes such as teaching practices. Given its importance for student learning, teaching quality needs to be evaluated using measures that are reliable and valid (Klieme 2013; Marsh and Roche 1997; Müller et al. 2016; Wallace et al. 2016).

This study evaluates indicators of science teaching quality provided by the 2015 cycle of the Programme for International Student Assessment (PISA, OECD 2016). We focus on PISA for several reasons. First, while many studies have used PISA's teaching scales to address substantive questions (Aditomo and Klieme 2020; Gee and Wong 2012; Hwang et al. 2018; Jiang and McComas 2015; McConney et al. 2014), few have focused on evaluating their psychometric properties. We found only one study which examined PISA's teaching scales across all participating countries (Wenger et al. 2018). Wenger et al.'s study did not consider teaching quality in science and only focused on reliability.

Second, PISA randomly samples students (as opposed to classes) from schools. Thus, students from the same school likely come from different classes and may be taught by different teachers. When aggregated at the school level, student ratings in PISA do not refer to specific teachers (Wang and Degol 2016). For this reason, studies utilizing PISA data have mostly avoided the aggregation of its teaching measures by treating them as individual-level constructs (Aditomo and Klieme 2020). This is problematic because teaching is mostly a class-level process, and hence, its effect is best assessed at the classroom level. Some authors have conceptualised teaching quality as a dimension of school climate, suggesting the possibility of assessing it at the school level (Samuel n.d.; Wang and Degol 2016). Few studies to date have critically addressed the extent to which PISA's teaching measures can be used to assess teaching as a school-level characteristic.

1.1 Conceptualising teaching quality

Teachers and the way they teach are major factors which determine students' learning outcomes. Teachers vary in how effective they are in improving cognitive outcomes (Blazar 2015; Gershenson 2016; Jackson 2012; Slater et al. 2012) as well as affective and behavioural outcomes (Jennings and Di Prete 2010; Kraft and Grace 2016). Similarly, there is also a large variation in the effects of different teaching practices on student learning, as powerfully shown in the work of John Hattie (2008; 2017). Teacher and teaching effects are linked in that effective teachers are those who practise more effective teaching. One study indicated that the teacher effects seem to be less related with their background, experience, and qualifications, and more with what they do in their classroom, i.e. teaching practices (Slater et al. 2012). Another study showed that specific teaching practices explain why some teachers are more effective than others, with emotional support predicting differences in affective outcomes and classroom management predicting behavioural outcomes (Blazar and Kraft 2017).

Teaching quality can be described in terms of generic dimensions which characterise effective teaching practices. A number of frameworks have been proposed to describe these key dimensions. According to one framework, teaching quality has three basic dimensions: classroom management, student support, and cognitive activation (Klieme et al. 2009; Praetorius et al. 2018). Each of these dimensions facilitates different aspects of learning, namely time on task, motivation, and knowledge construction. As detailed below, the relations between each dimension with learning are supported by different theories and research traditions.

Classroom management refers to the organisation and structure of lessons which involves the establishment of clear rules, frequent monitoring of student behaviour, effective response to disruptions, and efficient use of time (Praetorius et al. 2018). The importance of classroom management is highlighted in early models of school learning which defined learning opportunity in terms of "time on task" (Carroll 1989). In that model, good classroom management produces an orderly climate which allows students to focus their attention to the relevant materials and activities. The opportunity to learn provided by an orderly climate may not necessarily lead students to develop a deeper understanding of the materials. However, a disruptive climate is assumed to lead to frustration and diminished motivation (Carroll 1989; Egeberg et al. 2016; Emmer and Stough 2001). In other words, good classroom management should correlate positively with student motivation. In many studies, including PISA, classroom management is measured by way of students' report of how orderly or disruptive their typical lessons are.

Student support, the second teaching quality dimension, refers to teacher actions which cater to students' psychological needs. The importance of student support is emphasised by theories of motivation such as self-determination theory (Ryan and Deci 2000). According to this theory, teachers and schools should strive to fulfil students' basic psychological needs of autonomy (feeling empowered to exercise individual choice), belongingness (feelings of being valued members of a community), and competence (feelings of having the opportunity to learn and grow). In PISA, student support is most directly reflected in students' reports of emotional support, personal feedback, and adaptive instruction. From a self-determination theory perspective, these measures should help fulfil students' needs, which in turn should promote intrinsic motivation (Deci et al. 1991).

Two other teaching measures in PISA, teacher-directed and inquiry-based instructions, can also be seen as catering for students' psychological needs albeit in a less direct manner. Teacher-directed instruction provides cognitive scaffolds which, when properly implemented, should help students feel competent. Inquiry-based instruction, again when properly implemented, provides room for personal choices (e.g. in designing experiments and analysing data) which should help cater for students' need for autonomy. Hence, both forms of instruction should be correlated positively with intrinsic motivation.

The final basic teaching quality dimension, *cognitive activation*, refer to activities which prompt students to engage in deep processing of the learning materials. The importance of cognitive activation is highlighted by constructivist theories of learning, which assume that learning can only occur through active construction of knowledge

by the learner (Bransford et al. 2000; Derry 1996; Wittrock 2010). Cognitive activation can be implemented through the provision of tasks which activate relevant prior knowledge, the scaffolding metacognitive processes, and the use of questions and collaborative discourse around important ideas (Praetorius et al. 2018). Theoretically, measures of cognitive activation should predict students' scores in achievement tests. However, unlike in previous cycles of PISA, in 2015, cognitive activation was not directly assessed.

1.2 Teaching quality as a classroom-level phenomenon

Teaching is primarily a classroom process, and therefore, teaching quality is considered to be a classroom-level phenomenon (Cooley and Leinhardt 1980). Due to the comparative proximity of the classroom environment to students' experiences, classroom-level processes such as teaching are considered to bear stronger influence on student achievement compared with school-level factors (Kyriakides et al. 2000; Scheerens and Creemers 1989). Indeed, teaching quality is seen as a key factor which mediates the effects of tangible components of the education system (e.g. teacher qualifications, school infrastructure, and programmes) on student learning outcomes (Creemers 1994). Several studies have found that student achievement often varies more between classrooms than between different schools (Goldstein 1997; Muthe'n 1991). This further suggests that teaching quality may vary substantially between classrooms within the same school.

The implication is that teaching quality should be evaluated at the classroom level. Analysis which ignores the classroom level has been shown to produce distorted estimates of variance, biased standard errors, and thus potentially misleading conclusions about the relationships between the variables of interest (Hutchinson and Healy 2001; Moerbeek 2004). More specifically, omitting the classroom level in such analysis tends to inflate the between-school variance, while simultaneously underestimating the overall effect of schooling on student achievement (Martínez 2012). Furthermore, without explicitly modelling classroom-level factors, the strength of relationships between process variables and student achievement can be mistakenly attributed to school-level factors—thereby masking inequalities in educational opportunities which may exist within each school (Martínez 2012).

Overall, these considerations suggest that in addressing questions regarding the effects of schooling on student learning, and the mechanisms which explain those effects, researchers need to explicitly model classroom-level variables. This poses a particular problem for researchers who are working with data which omits classroom identification such as PISA. Can such data yield meaningful insights regarding the quality of teaching? We postpone discussing this issue after describing the use of student ratings to evaluate teaching.

1.3 Using student ratings to evaluate teaching quality

Many studies evaluate teaching quality based on student rating data. In addition to being relatively efficient, such data are typically based on students' extensive experience of the assessed behaviours. Moreover, student ratings reflect students'

interpretations of the learning environment, which is an important mediator between teaching and learning. Research suggests that student ratings are not simply a function of teacher popularity (Fauth et al. 2014; Kunter and Baumert 2006) and can yield insights that complement information from teacher self-reports (Aldrup et al. 2018; Kunter and Baumert 2006).

Regardless of whether teaching is analysed at the classroom or school, evaluating teaching quality based on student ratings requires the aggregation of individual-level (L1) data into the group-level (L2). The current methodological best practice to achieve this is through the application of doubly-latent multilevel models (Morin et al. 2013). In such models, multiple ratings from each student are treated as manifest (observable) indicators of a latent (unobservable) construct. In this case, teaching quality is regarded as a construct that is latent in relations to multiple items in a scale measuring a specific aspect of teaching. In other words, the construct represents each student's perception of an objective aspect of teaching. Inconsistencies between item responses are regarded as measurement error.

In addition, these models are also latent in the sense that L2 constructs are inferred based on latent aggregations of responses from multiple students at L1 (Morin et al. 2013). In other words, teaching quality at L2 is inferred from latent L1 constructs, instead of being simply formed through averaging manifest responses. Here students from the same classroom may differ in their rating of the teaching quality. This variation between latent L1 constructs can be interpreted as either representing sampling error or meaningful variation in how an aspect of teaching is implemented within the same lesson/classroom. This latter interpretation makes sense when teachers differentiate their teaching in response to individual students' needs within the same lesson.

Doubly-latent multilevel models guard researchers from the trap of ecological fallacy, i.e. falsely assuming that effects observed at one level can be generalised to another level (Morin et al. 2013). In addition, such models also allow researchers to simultaneously control for measurement error due to the sampling of items, as well as sampling error due to sampling of students from a classroom or school (Morin et al. 2013). Measurement error is controlled for by utilizing multiple manifest indicators to measure a latent construct, while sampling error is controlled through the aggregation of multiple students to represent classroom or school-level constructs. As an illustration, a doubly-latent multilevel model for teacher-directed instruction is depicted in Fig. 1.

1.4 Reliability and validity of teaching quality measures

The use of doubly-latent multilevel models to assess teaching quality means that reliability and validity need to be ensured at both L1 and L2. Reliability is typically assessed by calculating the internal consistency, which can be estimated at both levels (Hox 2010; Miller and Murdock 2007; Raudensbush and Bryk 2002). To determine internal consistency at L1, the most commonly applied measures are Cronbach's alpha and different forms of the r_{WG} index proposed by James et al. (1984); e.g. Wenger et al. 2018). The L1 reliability of the PISA teaching scales has been examined and found to be adequate (OECD 2017).

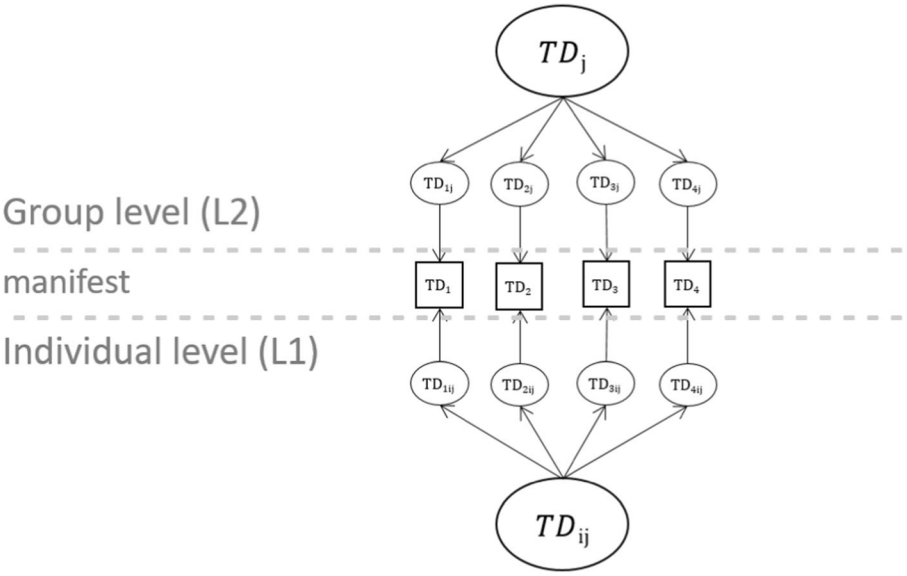

Fig. 1 Measurement models representing teacher-directed instruction

Meanwhile, internal consistency at L2 is often referred to as ICC-1 (Shrout and Fleiss 1979) and is defined by:

$$\mathrm{ICC{-}1} = \frac{\sigma^2_{\eta_B}}{\sigma^2_{\eta_B} + \sigma^2_{\eta_W}}$$

where $\sigma^2_{\eta_W}$ is the variance of the latent trait η at the within level (L1) and $\sigma^2_{\eta_B}$ is the variance of the latent trait η at the between level (L2). It hence informs about the proportion of variance at L2. ICC-1 can take values between 0 and 1. High values indicate that a lot of variance in the latent variable is due to the clustering of individuals.

Another essential measure for evaluating reliability of latent constructs at L2 is ICC-2, which is calculated similarly as ICC-1 (Raudenbush and Bryk 2002; Shrout and Fleiss 1979):

$$\mathrm{ICC{-}2} = \frac{\sigma^2_{\eta_B}}{\sigma^2_{\eta_B} + \frac{\sigma^2_{\eta_W}}{\overline{n}}}$$

where $\overline{n}$ represents the average cluster size. Dividing the variance of the latent trait η at the within level by the average cluster size has the effect that for large cluster sizes, the proportion of within-level variance in the denominator will be small, which will result in increased ICC-2 values. ICC-2 is hence considered the reliability of group mean scores in relation to sampling error (Lüdtke et al. 2011; Stapleton et al. 2016). Sampling error occurs when not all students in a class or school provide data. Like ICC-1, ICC-2 can take values between 0 and 1.

Multilevel models also need to be checked in terms of their validity. One important source of evidence to consider is the unidimensionality of the model at both L1 and L2. This is depicted visually in Fig. 1 for the case of teacher-directed instruction. This model assumes that the inter-item covariation can be explained by one latent factor. At L1, this means that a student's response to one item should be consistent with responses to the other items because the student has a certain perception of teacher-directed instruction in his/her science lessons. Correspondingly at L2, the aggregated response to an item (from students in a school) should be consistent with the aggregated responses to the other items because the students have a shared perception about teacher-directed instruction in their school. The extent to which these assumptions are met is the issue of structural or factorial validity which can be evaluated using multilevel confirmatory factor analysis (Brown 2015).

Model fit within the confirmatory factor analysis (CFA) framework can be evaluated via several indices. The standardised root mean square residual (SRMR) conceptually reflects the discrepancy between observed and model-predicted correlations. The SRMR is particularly useful for evaluating multilevel models because it can be calculated separately for L1 and L2. In this paper, we focus on the SRMR at the between level, since we are particularly interested in how well the model describes the variance covariance matrix at the school level. Another index is the root mean square error approximation (RMSEA), which is based on the chi-square but takes into account sample size and model complexity (where complex models are penalised and more parsimonious ones are rewarded). Smaller RMSEA values reflect better models (Hu and Bentler 1999). Other fit indices, such as the Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI), are based on the comparison of the proposed model chi-square to that of a more restricted baseline model. Typically, this baseline is an "independence model" which assumes that the covariances among items are zero.

In addition to factor structure, another source of validity evidence is the relation between the construct of interest and other relevant variables. Since our study concerns teaching effects, we are interested in whether teaching quality is related to learning outcomes. As described in a preceding section, the teaching measures in PISA are theoretically related to intrinsic motivation, such that schools with better teaching quality should also have students with higher collective intrinsic motivation. The relationships between intrinsic motivation and classroom management, student support, and teacher-directed and inquiry-based instructions have been found in prior empirical studies (Aditomo 2020; Aditomo and Klieme 2020; Aldrup et al. 2018; Decristan et al. 2016; Fauth et al. 2014; Kunter and Baumert 2006; Rjosk et al. 2014; Schiepe-tiska 2019; Wallace et al. 2016). Thus, associations with intrinsic motivation at the school level can be used as one source of evidence to assess the validity of teaching quality measures in PISA.

1.5 Plausibility of evaluating school-level teaching quality

As discussed above, teaching is primarily a classroom process and hence its quality should be evaluated at the classroom level. However, PISA randomly selects students from within a school. Therefore, the L2 in PISA refers not to intact classes, but to

schools which may be composed of more than one class and one teacher. This raises concerns regarding the meaning, reliability, and validity of teaching quality when measured at the school level.

Nonetheless, there are reasons to suggest the plausibility of interpreting PISA's teaching quality scales at the school level. The simplest reason is that students in a school may be referring to the same science teacher, even when they are not from the same classroom. This possibility is stronger in schools that are relatively small, and in schools or education systems where science is taught as a general science subject (and hence could be taught by the same teacher). If this is the case, then school-level teaching quality actually reflects teacher effects. Unfortunately, PISA does not collect information to identify schools which employ only one science teacher.

Even if students in a school are referring to different teachers, there are other reasons to support interpreting teaching quality as a school-level property. Theoretically, teaching quality can be seen as a part of school climate. School climate has been defined as perceptions about values, beliefs, interactions, and relationships held by students and staff within a school (Rudasill et al. 2018). School climate is often described as covering three domains: academic, social, and organisational (Wang and Degol 2016).

From this perspective, teachers in a school may have shared values and beliefs regarding what constitute good teaching, i.e. the kinds of teaching practices and student-teacher interactions which are encouraged and supported (e.g. via training programmes) by the school leadership and community. For example, some school policies may encourage teachers to be more attentive to students' individual learning needs (and hence allowing teachers to implement higher levels of adaptive instruction, emotional support, and personal feedback). Other schools may have whole-school positive discipline programmes, which are implemented by teachers in the form of more effective classroom management strategies.

A more technical reason has to do the wording of the items in the PISA teaching scales. Items in most of the scales refer to some aspect of the classroom situation or learning environment: teachers, teacher behaviour, teaching practice, or classroom climate. For example, classroom management items refer to students as a group ("Students don't listen to what the teacher says") or the classroom situation ("There is noise and disorder"). Adaptive instruction items refer to the teacher's typical behaviour (e.g. "The teacher adapts the lesson to my class's needs and knowledge"). Teacher-directed instruction items refer to classroom activities (e.g. "A whole class discussion takes place with the teacher"). The one exception is the personal feedback scale, which is composed of items referring to experience of the individual student. The item "*The teacher* tells *me* how I am performing in this course", for instance, reflects the student's personal experience of obtaining feedback. Because students may vary in their experience of feedback, the L1 latent factor bears substantive meaning. From this perspective, the personal feedback scale should exhibit relatively lower L2 reliability and validity compared with the other teaching scales.

1.6 Research questions

The preceding section presents reasons for the plausibility of interpreting PISA's teaching quality scales at the school level. The extent to which they are supported by

empirical data is the focus of this article. This is addressed by examining six scales provided by PISA 2015 intended to assess the following dimensions of teaching quality in science: classroom discipline, adaptive instruction, emotional support, constructive feedback, teacher-directed instruction, and inquiry-based instruction.

We are aware of only one prior study that investigated the reliability of PISA's school-level teaching quality indicators (Wenger et al. 2018). That study focused on mathematics and reading. Our study contributes to the literature by considering not only reliability but also factorial and predictive validity of the teaching scales. We focus on the science domain in PISA 2015. Our specific research questions are as follows:

1. How reliable are PISA's science teaching scales when used to assess school-level teaching quality, and how much does reliability vary across regions/countries?
2. To what extent do the teaching scales fit a two-level unidimensional factor structure, and how much does this vary across regions/countries?
3. Do the school-level factor scores exhibit positive relations with intrinsic motivation, and how much does this vary across regions/countries?

Regarding research question three, we postulate that higher ratings of teaching quality should be associated with higher/better intrinsic motivation to learn. This postulation is based on theoretical considerations outlined in the preceding section.

2 Method

2.1 Sample

The PISA student sample is drawn from a two-stage stratified procedure in which each participating country/region randomly selects schools, and then randomly selects 15-year-old students in those schools. We use the PISA 2015 sample which consisted of 503,146 students from 17,678 schools in 69 countries/regions (see Appendix for details). We include all regions or countries in PISA for practical reasons. PISA's strength is in the international nature of its database, and hence, researchers often use PISA to conduct comparative analysis across many different regions/countries. We intend our current study to be useful to future researchers who wish to examine science teaching quality using PISA 2015, in whichever region/country they are interested.

2.2 Instruments

The six teaching scales in PISA 2015 are (OECD 2017):

- *Adaptive instruction*—three items referring to the science teacher's adaptations of his/her instruction in response to the students' needs (e.g. "The teacher adapts the lesson to my class's needs and knowledge" and "The teacher changes the structure of the lesson on a topic that most students find difficult to understand").
- *Classroom management*—five items describing the disciplinary climate in the science lessons (e.g. "Students don't listen to what the teacher says" and "There is noise and disorder").

- *Teacher-directed instruction*—four items referring to teacher-led activities in the science lessons (e.g. "The teacher explains scientific ideas" and "A whole class discussion takes place with the teacher").
- *Emotional support*—five items describing the teacher's commitment to create a supportive climate in the science lessons (e.g. "The teacher shows an interest in every student's learning" and "The teacher continues teaching until the students understand").
- *Personal feedback*—five items describing personal feedback that the student receives from the science teacher (e.g. "The teacher tells me how I am performing in this course" and "The teacher tells me in which areas I can still improve").
- *Inquiry-based instruction*–eight items describing the use of inquiry activities in the science lessons (e.g. "Students spend time in the laboratory doing practical experiments" and "Students are required to argue about science questions").

Meanwhile, intrinsic motivation is defined as enjoyment and interest in learning about science. It is measured using 5 items (e.g. "I generally have fun when I am learning <broad science> topics"). The four response options for each item on every scale were *in all lessons* (or *every lesson*), *most lessons*, *some lessons*, and *never or hardly ever*. For this study, all items were scored such that higher scores reflected higher levels of teaching quality.

2.3 Analysis

Data preparation was conducted using SPSS v.23, while all analyses were conducted using Mplus 8 (Muthén and Muthén 2017). To answer research question one, reliability at L2 is determined by calculating the ICC-1 and ICC-2 for each scale in each participating country/region. The within- and between-school variances of the latent factors estimated from the two-level CFA—with students at L1 and schools at L2—were used to compute ICC-1 and ICC-2 according to Eqs. 1 and 2, respectively.

For shared constructs, ICC-1 values often lie around 0.10 and seldom exceed 0.30 (see, e.g., Klein et al. 2000; Stapleton and Hancock 2016; Wagner et al. 2016). Meanwhile, LeBreton and Senter (2008) suggest that values of 0.01, 0.10, and 0.25 correspond to small, medium, and large ICC-1 values. For manifest measures, Klein et al. (2000) provide rules of thumb for ICC-2: 0.7 is considered acceptable; 0.5 is considered marginally reliable. LeBreton and Senter (2008) recommend values between 0.70 and 0.85 to justify aggregation of ratings. Note, however, that ICC-2 evidently depends on the number of clusters and the inter-correlation of the individuals. This means that ICC-2 values rise with increasing average cluster sizes.

Regarding our second research question, the fit of the two-level CFA models is examined. We provide the general model fit indices CFI and RMSEA, but more importantly the SRMR at the between level. For interpretation purposes, SRMR values close to or below 0.08 are considered to reflect good fit (Hu and Bentler 1999). Smaller RMSEA values reflect better models, with values close to 0.06 or below considered acceptable (Hu and Bentler 1999). For the CFI and TLI, values above 0.95 indicate good fit, and values between 0.90 and 0.95 indicate marginal fit (Brown 2015; Hu and Bentler 1999).

All scales are assumed to have latent factors at both the individual and school levels. For those CFAs, we did not assume invariance, meaning that factor loadings at the student and school levels are assumed to be equivalent (Schweig 2014; Zyphur et al. 2008), because we did not necessarily assume that the meaning of the construct at L1 was the same at L2. However, in five countries, namely Canada, Chile, China, Colombia, and Costa Rica, we tested cross-level invariance of the scales in order to check whether this drastically changes model fit indices.

To answer our third research question, the latent factor representing school-level intrinsic motivation (measured by 5 items) was regressed on the latent school-level teaching quality factors for each scale and each region/country. Motivation is chosen as the external criterion because there is strong theoretical rationale, as well as consistent prior empirical findings, to expect positive associations with the teaching scales included in PISA 2015. All the CFA and SEM used the robust maximum likelihood estimator.

3 Results

3.1 Reliability

Table 1 summarises the results of the ICC-1 and ICC-2 values across all countries/regions for each of the scales. As is evident from the table, reliability differs across the scales. In most countries/regions, school-level reliability is high for the classroom management scale and low for the inquiry and adaptive instruction scales. The other scales could be regarded as on average being marginally reliable at the school level. Note that these results also indicate large variability across countries/regions (depicted visually in Fig. 1). Hence, in certain countries/regions, the classroom management scale can be unreliable, and the inquiry and adaptive instruction scales can have adequate reliability. Researchers who aim to work with the data at the school level are referred to the Appendix for exact values for each country/region. Comparing the empirical values with the recommended values, we provided in Section 1, they can make a decision of whether the reliabilities in that particular country are sufficient.

3.2 Internal structure

On average, the classroom management, emotional support, and feedback scales exhibited good SRMR between values in almost all countries/regions. In contrast, the inquiry and the teacher-directed instruction scales exhibited poor fit in most countries/regions. We paid particular attention to SRMR between values, since they report how well the model fits at the school level. The remaining fit indices, which largely reflect the model fit at L1, support the findings of model fit at L2. Only the teacher-directed instruction scale shows slightly better fit at L1, especially regarding the CFI, compared with fit at L2. Note that fit indices for the adaptive instruction scale cannot be meaningfully interpreted, because with only three items, the measurement model is saturated. Again, as depicted in Fig. 2, the fit indices of each scale vary across different countries/regions.

Table 1 School-level reliability indices (summary across all regions/countries)

Scale	Index	Mean	SD	Median	Min	Max
Adaptive instruction	ICC-1	0.06	0.02	0.06	0.01	0.11
	ICC-2	0.58	0.12	0.60	0.17	0.82
Classroom management	ICC-1	0.12	0.05	0.11	0.04	0.29
	ICC-2	0.76	0.09	0.76	0.38	0.93
Teacher-directed instruction	ICC-1	0.08	0.06	0.06	0.01	0.38
	ICC-2	0.64	0.16	0.64	0.14	0.91
Emotional support	ICC-1	0.08	0.03	0.07	0.00	0.17
	ICC-2	0.65	0.14	0.67	0.12	0.91
Personal feedback	ICC-1	0.07	0.04	0.06	0.00	0.23
	ICC-2	0.62	0.15	0.63	0.00	0.91
Inquiry-based instruction	ICC-1	0.05	0.07	0.02	0.00	0.39
	ICC-2	0.41	0.29	0.39	0.00	0.94

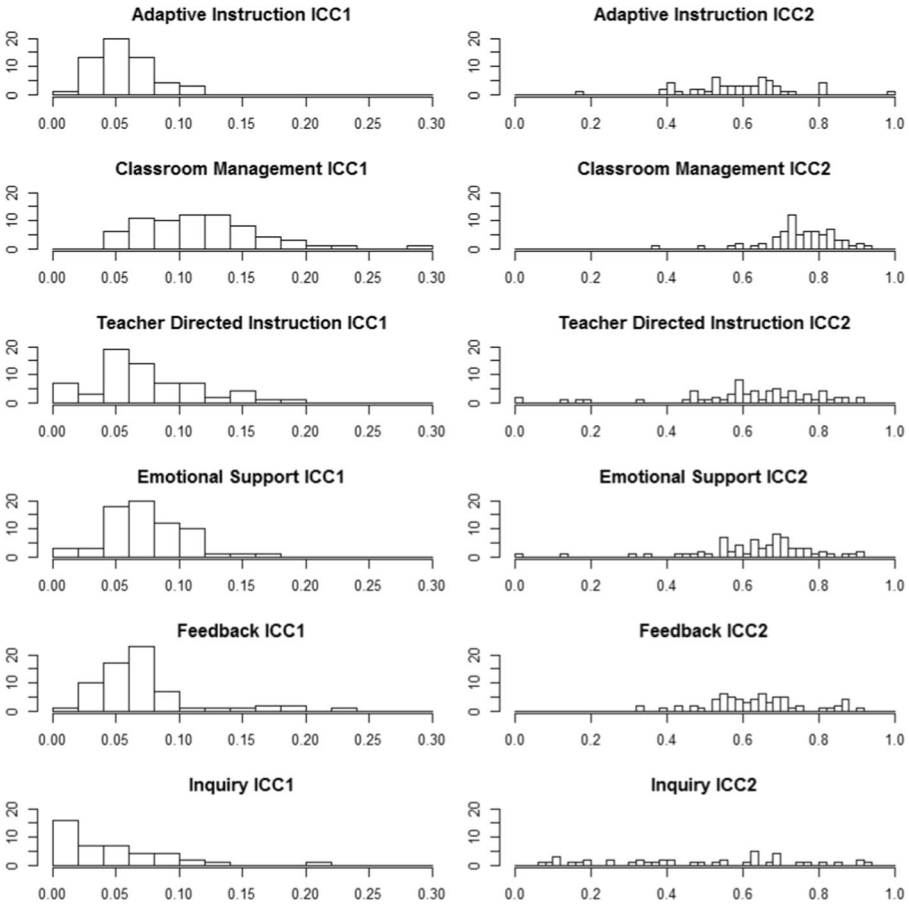

Fig. 2 Distribution of ICC-1 and ICC-2 values across all countries for science teaching scales in PISA 2015

The comparison between the models with and without invariance constraints across the levels, which we conducted in five countries, showed inconclusive results regarding which model is preferable according to the AIC and BIC. The change in the CFI was very small for all scales in almost all countries, and the RMSEA values hardly differed (Figure 3, Table 2).

3.3 Relationship with intrinsic motivation

On average, higher scores of adaptive instruction, classroom management, teacher-directed instruction, and emotional support predicted higher levels of intrinsic motivation to learn science. Meanwhile, the average regression coefficients for the feedback and inquiry scales were close to zero and mostly non-significant (Table 3). Again, there is considerable variation across countries/regions in the magnitude and statistical significance of the regression coefficients (see the Appendix for country-specific values).

4 Discussion

This study examined the extent to which student ratings from PISA provide reliable and valid information about teaching quality at the school-level. With a focus on science in

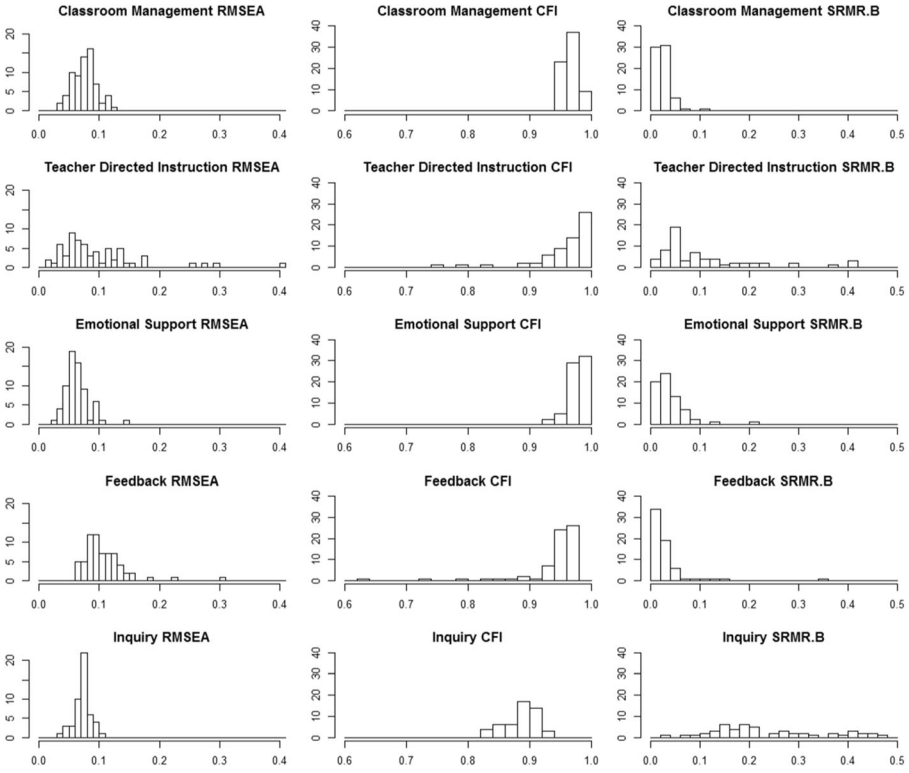

Fig. 3 Fit of two-level unidimensional measurement models of the science teaching scales in PISA 2015

Table 2 Fit of measurement models (summary across all regions/countries)

Scale	Fit index	Mean	SD	Median	Min	Max
Adaptive instruction	CFI	1.00	0.00	1.00	1.00	1.00
	RMSEA	0.00	0.00	0.00	0.00	0.00
	SRMR between	0.00	0.00	0.00	0.00	0.01
Classroom management	CFI	0.97	0.01	0.97	0.94	0.99
	RMSEA	0.08	0.02	0.08	0.03	0.12
	SRMR between	0.03	0.02	0.02	0.01	0.11
Teacher-directed instruction	CFI	0.93	0.16	0.97	0.00	1.00
	RMSEA	0.12	0.17	0.08	0.01	1.34
	SRMR between	0.11	0.10	0.07	0.01	0.42
Emotional support	CFI	0.98	0.01	0.98	0.93	1.00
	RMSEA	0.06	0.02	0.06	0.03	0.14
	SRMR between	0.04	0.03	0.03	0.00	0.21
Personal feedback	CFI	0.94	0.06	0.96	0.62	0.98
	RMSEA	0.11	0.04	0.10	0.06	0.31
	SRMR between	0.05	0.10	0.02	0.01	0.76
Inquiry-based instruction	CFI	0.89	0.03	0.89	0.82	0.93
	RMSEA	0.07	0.01	0.07	0.04	0.11
	SRMR between	0.24	0.11	0.20	0.03	0.47

PISA 2015, we calculated ICC-1 and ICC-2 to investigate school-level reliability of the two scales measuring generic quality, namely classroom management and emotional support, and the four instructional practices scales, namely inquiry-based instruction, individual feedback, adaptive instruction, and teacher-directed instruction. We also examined the factorial and predictive validity of those six scales.

A potential problem with regard to assessing teaching quality in PISA arises from its sampling procedure. Because students are sampled randomly from schools, the PISA sample lacks a teacher/classroom level (OECD 2017). Instead, the second level (L2) in PISA directly comprises the school. Thus, responses from students within the same school may pertain to different teachers. This raises the question of whether student ratings in PISA, which reflect student perceptions of different teachers, can be

Table 3 Average standardised regression coefficients across all regions/countries reflecting the latent school-level (L2) relationships between teaching quality and intrinsic motivation

Scale	Mean	SD	Median	Min	Max
Adaptive instruction	0.55	0.36	0.60	− 0.86	1.21
Classroom management	0.58	0.20	0.63	− 0.06	0.90
Teacher-directed instruction	0.69	0.28	0.74	− 0.60	1.02
Emotional support	0.30	0.39	0.37	− 0.69	0.90
Personal feedback	− 0.06	0.50	− 0.09	− 0.93	1.03
Inquiry-based instruction	− 0.08	0.51	− 0.13	− 1.05	0.87

aggregated at the school level. Our analyses show that the answer depends on the specific scale and country/region.

For the classroom management scale, L2 reliability was fairly high with an average ICC-1 of 0.116 and ICC-2 of 0.755. Using the cutoff points of 0.05 for ICC-1 and 0.70 for ICC-2, this scale could be judged as a sufficiently reliable measure of L2 teaching quality in the vast majority of the countries/regions examined. This means that within a school, students tend to agree on the organisation and structure of science lessons. The measurement model for this scale also showed a good fit in most countries/regions. Overall, these findings indicate that the scale can be used to reliably measure a unidimensional latent factor at both the individual and school levels.

Unfortunately, the same conclusion cannot be made about the other science teaching scales in PISA 2015. On average, the ICC-1 values for emotional support, individual feedback, and teacher-directed instruction scales were around 0.07 and 0.08. This represents a non-trivial agreement between students in a school (Lebreton and Senter 2008), suggesting that to some extent, these dimensions of teaching are class-spanning and can still be considered a feature of the school. For the inquiry and adaptive instruction scales, ICC-1 values were lower (about 0.05) but still indicate some level of agreement between students in a school. However, given the cluster size in the sample (which was around 30 students per school), the ICC-2 for these scales did not reach 0.70 in most countries/countries.

With regard to their factorial validity, the teacher-directed instruction, emotional support, and feedback scales exhibited good fit with the data in most countries/regions. This supports the use of these scales to measure a single latent factor of teaching quality at the student and school levels (compare with, e.g., Wagner et al. 2013). Meanwhile, the measurement model of the inquiry scale exhibited poor fit with the data in almost all countries/regions. This finding is consistent with prior research, which applied a more exploratory approach and found that the inquiry scale is not unidimensional (Aditomo and Klieme 2020; Lau and Lam 2017). Rather, the scale seems to tap into guided and unguided forms of science inquiry.

In general, the school-level reliabilities observed in this study are low when compared with those reported by studies which utilize class or teacher-based—as opposed to school-based—student samples to assess teaching (Fauth et al. 2014; Lüdtke et al. 2009; Lüdtke et al. 2006). On the one hand, this is not surprising given that in the PISA sample, students in the same school report their perceptions about different teachers. Thus, lower reliability estimates are observed because the aggregated scores reflect not only subjective individual experiences of the same learning environment but also objectively different targets of perception (i.e. different classrooms and teachers). On the other hand, the low reliabilities raise the question of what the PISA teaching scales actually measure at the school level. Put differently, how meaningful is it to assume the existence of constructs which reflect school-level teaching quality?

In this respect, perhaps it is not coincidental that the most reliable scale in PISA 2015, that is, classroom management, does not directly assess teaching. Instead of measuring teacher activities or behaviours, items of this scale refer to disciplinary climate. For example, the scale asks students to consider how often lessons are

disrupted due to noisy or unruly student behaviour. This distinction is important because disciplinary climate is not only a function of teachers' classroom management skills but also of student-related factors such as the classroom SES and prior achievement composition, which are typically more reliable measures at L2. Meanwhile, items of the other teaching scales refer directly to teacher behaviour or teaching activities which may differ between teachers within the same school. This line of reasoning suggests that the classroom management scale is more reliable at L2 because it does not directly measure teaching.

Does this mean that the other PISA teaching scales cannot be used to assess teaching quality at the school level? This is not necessarily the case. Our findings show that although the quality of science teaching varies within each school, there is also some meaningful variation of teaching quality between schools. Even if the level of ICC-1 is relatively low, it still lies above 0.05 for the inquiry scale and is often above 0.07 for the other scales. The caveat is that with relatively low ICC-1, one would need larger numbers of students per school to achieve adequate reliability. This can be illustrated by comparing Hungary and Montenegro, two countries with the same ICC-1 for adaptive instruction (about 0.056) but very different cluster sizes. In Hungary, where only 18 to19 students were sampled in each school, the ICC-2 was 0.520. In contrast, almost 75 students per school were sampled in Montenegro, which resulted in a much higher ICC-2 of 0.811 for the same scale. Thus, the adaptive instruction scale provides a reliable estimate of school-level teaching in Montenegro, but not in Hungary.

Results from our investigations regarding predictive validity also suggest that student ratings from some of the PISA scales capture meaningful differences of teaching quality between schools. Intrinsic motivation is a key affective outcome in science education, and there is strong theoretical ground proposing that teaching quality is positively associated with higher motivation (Deci et al. 1991; Deci et al. 1996; Ryan and Deci 2000). Our findings show that the classroom management, emotional support, teacher-directed instruction, and adaptive instruction scales were indeed predictive of higher intrinsic motivation in many countries/regions. This finding was not replicated for school-level feedback, which predicted higher intrinsic motivation in only 19% of the regions/countries for which data was available, and even predicted *lower* intrinsic motivation in some countries/regions. A reason for this might be that the feedback scale measures practices aimed at individual students. Unlike the other scales, the feedback scale is intended to measure a student-level construct. Thus, its effect at the aggregated level is better seen as reflecting a composition rather than a climate effect (Lüdtke et al. 2009). This conjecture is supported by additional analyses which show that at the student level, feedback was positively and significantly associated with intrinsic motivation in almost all the regions/countries. As theory suggests, within a school, students who were provided with more feedback were more likely to enjoy learning science.

Like the feedback scale, the inquiry scale failed to predict higher intrinsic motivation in most countries/regions. In this case, the most likely reason has to do with the scale's poor factorial validity. As mentioned, poor fit of the measurement model may be an indication that the scale is tapping onto a multidimensional construct. Indeed, previous

analysis of the inquiry scale for a sub-set of the PISA regions has shown that the scale reflects a two-factor structure representing teacher-guided and unguided forms of inquiry (Aditomo and Klieme 2020). When used to measure only a single latent factor, the inquiry scale yields a score that conflates the two dimensions. Considering the importance of guidance and structure in facilitating learning from inquiry (Lazonder and Harmsen 2016), such conflation likely masks the substantive relations between inquiry and learning outcomes.

5 Conclusions and implications

Our findings show that in most countries/regions, the teaching scales in PISA 2015 have low reliabilities when used to assess school-level teaching quality. The exception was for the classroom management scale, which suggests that effective classroom management is a school quality that spans across classes. For the other scales, adequate school-level reliability could only be observed in certain countries/regions. This stands in contrast to the high single-level reliabilities of these scales in all countries/regions reported by the OECD (2017, p. 313). Nonetheless, findings also indicate that the classroom management, emotional support, adaptive instruction, and teacher-directed instruction scales capture meaningful differences in teaching quality between schools. This opens the possibility of using student ratings of teaching in PISA to investigate school-level effects in certain countries/regions, thereby allowing extensions of prior analysis that thus far only utilized the data at the individual student level.

To reiterate, the relatively low aggregate reliabilities of PISA's teaching scales observed in many countries is likely because student ratings are aggregated at the school level as opposed to teachers/classrooms. A clear implication of these findings is that researchers and policy makers wishing to make inferences about school-level teaching quality using PISA data should proceed cautiously and check the level 2 reliability for the specific regions or school sectors they wish to evaluate. Relatedly, researchers need to formulate a clear theoretical reasoning about the substantive meaning of measures of teaching at the school level. In addition, the findings point to the importance of using a doubly-latent approach (multilevel structural equation modelling), which uses latent factors and latent aggregation (Morin et al. 2013). In this way, relationships between school-level teaching quality and other variables can be estimated while taking into account the unreliability in the measurement model (Kelloway 2015).

Our study also points to several avenues for further research. The first is related to the inquiry scale, whose measurement model exhibited poor model fit and even failed to converge in many countries/regions (see Appendix). Given the centrality of inquiry-based instruction in science education, its assessment based on student ratings requires a more in-depth analysis than can be provided by our study. Also, our study did not examine the comparability of the teaching scales across different regions/countries (He and van de Vijver 2013). Although international studies like PISA adopt rigorous procedures to minimise cross-cultural bias, measurement invariance of its scales cannot be assumed (for the case of teaching quality in mathematics, see Fischer et al. 2019). However, examining this issue for the whole set of participating regions/countries in PISA was beyond the scope of the current paper.

Appendix

Table 4 Sample size per country/region

PISA code	Country/region	Student	School
12	Algeria	5519	161
36	Australia	14,530	758
40	Austria	7007	269
56	Belgium	9651	288
76	Brazil	23,141	841
100	Bulgaria	5928	180
124	Canada	20,058	759
152	Chile	7053	227
158	Chinese Taipei	7708	214
170	Colombia	11,795	372
188	Costa Rica	6866	205
191	Croatia	5809	160
203	Czech Republic	6894	344
208	Denmark	7161	333
214	Dominican Republic	4740	194
233	Estonia	5587	206
246	Finland	5882	168
250	France	6108	252
268	Georgia	5316	262
276	Germany	6504	256
300	Greece	5532	211
344	Hong Kong	5359	138
348	Hungary	5658	245
352	Iceland	3371	124
360	Indonesia	6513	236
372	Ireland	5741	167
376	Israel	6598	173
380	Italy	11,583	474
392	Japan	6647	198
400	Jordan	7267	250
410	Korea	5581	168
411	Kosovo	4826	224
422	Lebanon	4546	270
428	Latvia	4869	250
440	Lithuania	6525	311
442	Luxembourg	5299	44
446	Macao	4476	45
470	Malta	3634	59

Table 4 (continued)

PISA code	Country/region	Student	School
484	Mexico	7568	275
498	Moldova	5325	229
499	Montenegro	5665	64
528	Netherlands	5385	187
554	New Zealand	4520	183
578	Norway	5456	229
604	Peru	6971	281
616	Poland	4478	169
620	Portugal	7325	246
634	Qatar	12,083	167
642	Romania	4876	182
643	Russian Federation	6036	210
702	Singapore	6115	177
703	Slovak Republic	6350	290
704	Vietnam	5826	188
705	Slovenia	6406	333
724	Spain	6736	201
752	Sweden	5458	202
756	Switzerland	5860	227
764	Thailand	8249	273
780	Trinidad and Tobago	4692	149
784	United Arab Emirates	14,167	473
788	Tunisia	5375	165
792	Turkey	5895	187
807	FYROM	5324	106
826	UK	14,157	550
840	USA	5712	177
858	Uruguay	6062	220
970	B-S-J-G (China)	9841	268
971	Spain regions	26,294	976
974	Argentina (Ciudad Autónoma de Buenos)	1657	58
	Total	503,146	17,678

To assist interpretation, we calculate the percentage of countries/regions in which each scale fulfil the minimum criteria of reliability and validity. Note that we adopt liberal cutoff scores for each index (e.g. 0.90 for CFI and 0.10 for RMSEA and SRMR). While methodologists have recommended more stringent criteria (e.g. 0.95 for CFI; 0.06 for RMSEA; and 0.08 for SRMR, see Hu and Bentler 1999), we recognise any cutoff scores which need to be adopted with caution by considering other features of a study such as sample size and model complexity, as well as substantive or theoretical considerations (Marsh et al. 2004; Perry et al. 2015). It is not our intention to advocate a certain set of cutoff fit criteria and recommend the reader to inspect indices for specific scales in each country in subsequent tables

Table 5 Percentage of regions/countries meeting minimum reliability and validity criteria

Scale	ICC-1 (>= 0.05) (%)	ICC-2 (>= 0.70) (%)	CFI (>= 0.90) (%)	RMSEA (<= 0.10) (%)	SRMR between (<= 0.10) (%)	Predictive (positive and $p < 0.05$) (%)
Adaptive instruction	51.9	11.1	100.0	100.0	100.0	63.5
Classroom management	97.1	82.6	100.0	89.9	98.6	86.2
Teacher-directed instruction	65.6	34.4	89.1	64.1	64.1	79.6
Emotional support	80.9	35.3	100.0	97.1	97.1	45.9
Personal feedback	69.7	25.8	87.9	51.5	92.4	19.0
Inquiry-based instruction	34.0	14.0	38.0	98.0	6.0	17.8

Table 6 Level 2 reliability (intraclass correlations 1 and 2) of the teaching scales in PISA 2015

PISA code	Country/region	Adaptive instruction		Classroom management		Teacher-directed instruction		Emotional support		Personal feedback		Inquiry-based instruction	
		ICC-1	ICC-2	ICC-1	ICC-2	ICC-1	ICC-2	ICC-1	ICC-2	ICC-1	ICC-2	ICC-1	ICC-2
12	Algeria	NA	NA	0.060	0.684	0.094	0.776	0.080	0.744	0.045	0.613	0.000	0.000
36	Australia	0.057	0.470	0.108	0.649	0.066	0.513	0.063	0.503	0.076	0.550	NC	NC
40	Austria	0.083	0.654	0.199	0.840	0.009	0.164	0.139	0.770	0.050	0.524	0.031	0.407
56	Belgium	0.023	0.415	0.085	0.736	0.124	0.809	0.079	0.716	0.183	0.869	0.015	0.312
76	Brazil	0.101	0.656	0.146	0.767	0.102	0.669	0.074	0.592	0.064	0.545	0.103	0.677
100	Bulgaria	0.039	0.523	0.086	0.729	0.018	0.339	0.065	0.661	0.028	0.437	0.043	0.557
124	Canada	0.068	0.630	0.118	0.758	0.088	0.693	0.089	0.696	0.047	0.535	NC	NC
152	Chile	0.058	0.622	0.183	0.860	0.070	0.670	0.078	0.698	0.067	0.659	0.005	0.114
158	Chinese Taipei	0.037	0.563	0.067	0.705	0.088	0.762	0.034	0.541	0.060	0.681	NC	NC
170	Colombia	0.050	0.604	0.086	0.735	0.066	0.675	0.063	0.663	0.051	0.611	0.070	0.690
188	Costa Rica	0.042	0.538	0.075	0.713	0.060	0.631	0.056	0.643	0.070	0.666	0.000	0.000
191	Croatia	0.039	0.560	0.082	0.739	0.068	0.696	0.045	0.597	0.075	0.716	0.021	0.401
203	Czech Republic	0.057	0.536	0.172	0.797	0.044	0.470	0.091	0.652	NC	NC	NC	NC
208	Denmark	0.091	0.655	0.120	0.734	0.044	0.469	0.057	0.547	0.075	0.606	0.080	0.626
214	Dominican Republic	0.058	0.549	0.139	0.775	0.065	0.587	0.054	0.544	0.065	0.586	0.029	0.385
233	Estonia	0.086	0.709	0.110	0.763	0.006	0.135	0.062	0.634	0.033	0.472	NC	NC
246	Finland	0.046	0.620	0.104	0.797	0.044	0.610	0.053	0.651	0.050	0.640	NC	NC
250	France	0.029	0.391	0.125	0.756	NC	NC	0.094	0.691	0.074	0.630	NC	NC
268	Georgia	NA	NA	0.121	0.733	0.086	0.648	0.096	0.680	0.024	0.332	0.100	0.689
276	Germany	0.039	0.418	0.046	0.485	0.136	0.739	0.051	0.493	0.036	0.397	0.046	0.483
300	Greece	0.068	0.644	0.138	0.801	0.084	0.696	0.084	0.697	0.068	0.643	0.013	0.252

Table 6 (continued)

PISA code	Country/region	Adaptive instruction		Classroom management		Teacher-directed instruction		Emotional support		Personal feedback		Inquiry-based instruction	
		ICC-1	ICC-2	ICC-1	ICC-2	ICC-1	ICC-2	ICC-1	ICC-2	ICC-1	ICC-2	ICC-1	ICC-2
344	Hong Kong	0.027	0.429	0.083	0.713	0.142	0.819	0.016	0.304	0.052	0.598	NC	NC
348	Hungary	0.056	0.520	0.104	0.689	0.117	0.711	0.047	0.478	0.064	0.556	0.000	0.000
352	Iceland	0.055	0.583	0.199	0.859	0.063	0.620	0.114	0.758	0.078	0.671	NC	NC
360	Indonesia	NA	NA	0.152	0.828	0.049	0.582	0.058	0.627	0.188	0.862	0.060	0.632
372	Ireland	0.021	0.403	0.077	0.720	0.043	0.584	0.053	0.634	0.077	0.719	0.065	0.681
376	Israel	0.042	0.600	0.071	0.727	NC	NC	0.156	0.864	0.163	0.868	0.096	0.779
380	Italy	0.049	0.530	0.176	0.826	NC	NC	0.113	0.737	0.125	0.758	NC	NC
392	Japan	0.111	0.801	0.233	0.908	0.034	0.533	0.099	0.781	0.074	0.721	NC	NC
400	Jordan	NA	NA	0.152	0.836	0.152	0.834	0.083	0.719	0.034	0.497	0.095	0.747
410	Korea	0.070	0.696	0.159	0.853	0.046	0.596	0.072	0.702	0.085	0.740	0.123	0.810
411	Kosovo	NA	NA	0.097	0.693	0.065	0.589	NC	NC	0.165	0.805	NC	NC
422	Lebanon	NA	NA	0.287	0.869	0.378	0.909	0.074	0.568	0.232	0.831	0.007	0.104
428	Latvia	0.011	0.165	0.122	0.715	0.044	0.452	0.116	0.704	0.084	0.623	0.013	0.186
440	Lithuania	0.075	0.606	0.131	0.743	0.012	0.191	0.028	0.357	0.064	0.567	NC	NC
442	Luxembourg	0.020	0.679	0.061	0.875	0.060	0.871	0.076	0.898	0.054	0.856	0.008	0.470
446	Macao	0.025	0.688	0.042	0.795	0.020	0.639	0.108	0.914	0.103	0.910	0.006	0.333
470	Malta	NA	NA	0.146	0.911	0.140	0.907	0.068	0.815	0.095	0.862	0.081	0.842
484	Mexico	0.048	0.553	0.099	0.736	0.056	0.594	0.064	0.630	0.087	0.702	0.000	0.000
498	Moldova	NA	NA	0.124	0.764	0.049	0.539	0.167	0.821	0.098	0.711	0.393	0.937
499	Montenegro	0.056	0.811	0.054	0.811	0.043	0.767	0.114	0.906	0.023	0.630	0.120	0.911
528	Netherlands	0.075	0.653	0.090	0.701	0.147	0.799	0.069	0.635	0.071	0.640	0.010	0.187

Table 6 (continued)

PISA code	Country/region	Adaptive instruction		Classroom management		Teacher-directed instruction		Emotional support		Personal feedback		Inquiry-based instruction	
		ICC-1	ICC-2	ICC-1	ICC-2	ICC-1	ICC-2	ICC-1	ICC-2	ICC-1	ICC-2	ICC-1	ICC-2
554	New Zealand	0.040	0.468	0.074	0.632	0.041	0.479	0.052	0.540	0.059	0.574	0.005	0.096
578	Norway	0.087	0.679	0.148	0.799	0.053	0.558	0.100	0.714	0.079	0.659	0.000	0.000
604	Peru	0.044	0.491	0.059	0.579	0.077	0.640	0.113	0.732	0.052	0.536	0.015	0.248
616	Poland	0.063	0.634	0.113	0.767	0.034	0.479	0.080	0.691	0.049	0.568	0.019	0.330
620	Portugal	0.030	0.398	0.061	0.583	NC	NC	0.044	0.497	0.061	0.581	NC	NC
634	Qatar	0.072	0.816	0.172	0.926	0.089	0.850	0.041	0.716	0.021	0.559	0.009	0.348
642	Romania	NA	NA	0.217	0.881	0.178	0.853	0.042	0.541	NC	NC	NC	NC
643	Russian Federation	0.042	0.527	0.121	0.782	0.035	0.486	0.083	0.701	0.047	0.560	0.007	0.148
702	Singapore	0.077	0.729	0.114	0.806	0.116	0.808	0.048	0.619	0.063	0.682	0.063	0.685
703	Slovak Republic	0.035	0.417	0.169	0.803	0.114	0.719	0.101	0.690	0.083	0.642	0.010	0.169
704	Vietnam	NA	NA	0.136	0.830	0.061	0.668	0.066	0.685	0.027	0.461	0.034	0.525
705	Slovenia	NA	NA	0.075	0.377	NA	NA	0.097	0.441	NA	NA	NC	NC
724	Spain	0.075	0.686	0.102	0.758	0.089	0.726	0.096	0.743	0.066	0.658	0.022	0.381
752	Sweden	0.079	0.678	0.160	0.829	0.050	0.566	0.077	0.676	0.043	0.528	0.045	0.541
756	Switzerland	0.065	0.594	0.106	0.725	0.113	0.731	0.061	0.581	0.053	0.541	NC	NC
764	Thailand	NA	NA	0.066	0.659	0.051	0.593	0.027	0.432	0.049	0.582	0.058	0.624
780	Trinidad and Tobago	NA	NA	0.156	0.848	0.043	0.567	0.057	0.640	0.040	0.550	NC	NC
784	United Arab Emirates	0.072	0.666	0.115	0.771	0.078	0.685	0.098	0.738	0.079	0.687	0.003	0.073
788	Tunisia	0.047	0.566	0.051	0.598	0.055	0.612	0.076	0.690	0.000	0.000	0.004	0.103
792	Turkey	0.060	0.647	0.123	0.801	0.045	0.573	0.041	0.552	0.025	0.422	NC	NC
807	FYROM	NA	NA	0.051	0.720	0.056	0.735	0.003	0.121	0.045	0.684	0.031	0.604

Table 6 (continued)

PISA code	Country/region	Adaptive instruction		Classroom management		Teacher-directed instruction		Emotional support		Personal feedback		Inquiry-based instruction	
		ICC-1	ICC-2	ICC-1	ICC-2	ICC-1	ICC-2	ICC-1	ICC-2	ICC-1	ICC-2	ICC-1	ICC-2
826	UK	0.046	0.523	0.105	0.731	0.116	0.750	0.052	0.559	0.053	0.562	0.000	0.000
840	USA	0.044	0.571	0.077	0.710	0.050	0.603	0.043	0.569	0.069	0.683	0.054	0.623
858	Uruguay	0.045	0.495	0.090	0.690	0.064	0.594	0.061	0.586	0.023	0.328	0.000	0.000
970	B-S-J-G (China)	0.106	0.802	0.117	0.820	0.109	0.807	0.113	0.813	0.147	0.855	0.220	0.906
971	Spain (regions)	0.072	0.676	0.097	0.746	0.076	0.690	0.104	0.760	0.070	0.669	0.057	0.624
974	Argentina (Ciudad Autónoma de Buenos)	NA	NA	0.124	0.799	0.189	0.867	0.111	0.777	0.082	0.713	0.020	0.368

NA; not applicable due to unavailable data; *NC*, data available but model estimation failed to converge

Table 7 Fit indices of the two-level models of teaching scales in PISA 2015

PISA code	Region/country	Classroom management			Teacher-directed instruction			Emotional support			Feedback			Inquiry-based instruction		
		CFI	RMSEA	SRMR between	CFI	RMSEA	SRMR between	CFI	RMSEA	SRMR between	CFI	RMSEA	SRMR between	CFI	RMSEA	SRMR between
12	Algeria	0.976	0.044	0.058	0.988	0.057	0.022	0.978	0.052	0.069	0.919	0.086	0.126	0.915	0.046	0.278
36	Australia	0.962	0.100	0.008	0.992	0.054	0.033	0.981	0.075	0.019	0.966	0.099	0.010	NC	NC	NC
40	Austria	0.982	0.066	0.012	0.000	1.343	0.300	0.979	0.062	0.058	0.980	0.067	0.016	0.886	0.077	0.202
56	Belgium	0.977	0.067	0.022	0.975	0.064	0.016	0.983	0.055	0.043	0.950	0.095	0.029	0.882	0.072	0.291
76	Brazil	0.982	0.056	0.017	0.969	0.092	0.040	0.990	0.044	0.014	0.953	0.091	0.152	0.888	0.073	0.155
100	Bulgaria	0.970	0.079	0.018	0.960	0.117	0.115	0.980	0.059	0.021	0.961	0.092	0.027	0.888	0.084	0.176
124	Canada	0.965	0.086	0.018	0.985	0.069	0.052	0.969	0.086	0.020	0.958	0.097	0.035	NC	NC	NC
152	Chile	0.975	0.071	0.010	0.965	0.100	0.051	0.984	0.057	0.011	0.952	0.111	0.016	0.896	0.078	0.302
158	Chinese Taipei	0.975	0.074	0.008	0.942	0.137	0.124	0.967	0.095	0.040	0.935	0.133	0.032	NC	NC	NC
170	Colombia	0.976	0.056	0.026	0.934	0.114	0.046	0.989	0.043	0.017	0.975	0.073	0.020	0.866	0.073	0.187
188	Costa Rica	0.977	0.057	0.028	0.251	0.406	0.183	0.987	0.049	0.028	0.964	0.094	0.024	0.859	0.087	0.379
191	Croatia	0.966	0.087	0.011	0.753	0.291	0.203	0.990	0.044	0.029	0.950	0.111	0.009	0.902	0.076	0.271
203	Czech Republic	0.973	0.075	0.014	0.949	0.112	0.370	0.987	0.046	0.028	NA	NA	NA	NC	NC	NC
208	Denmark	0.955	0.095	0.018	0.942	0.132	0.123	0.986	0.050	0.014	0.937	0.124	0.014	0.884	0.069	0.142
214	Dominican Republic	0.953	0.083	0.039	0.985	0.064	0.042	0.997	0.025	0.037	0.961	0.088	0.031	0.900	0.063	0.175
233	Estonia	0.960	0.096	0.024	0.985	0.060	0.095	0.978	0.068	0.016	0.950	0.105	0.039	NC	NC	NC
246	Finland	0.977	0.076	0.011	0.970	0.096	0.050	0.979	0.066	0.028	0.960	0.105	0.013	NC	NC	NC
250	France	0.971	0.085	0.013	NA	NA	NA	0.981	0.063	0.023	0.938	0.125	0.058	NC	NC	NC
268	Georgia	0.950	0.079	0.043	0.976	0.065	0.066	0.980	0.049	0.034	0.970	0.070	0.021	0.901	0.057	0.109
276	Germany	0.987	0.051	0.023	0.919	0.155	0.089	0.983	0.062	0.055	0.978	0.069	0.024	0.897	0.072	0.209

Table 7 (continued)

PISA code	Region/country	Classroom management			Teacher-directed instruction			Emotional support			Feedback			Inquiry-based instruction		
		CFI	RMSEA	SRMR between	CFI	RMSEA	SRMR between	CFI	RMSEA	SRMR between	CFI	RMSEA	SRMR between	CFI	RMSEA	SRMR between
300	Greece	0.969	0.064	0.028	0.971	0.101	0.046	0.984	0.063	0.035	0.970	0.081	0.014	0.897	0.073	0.156
344	Hong Kong	0.966	0.089	0.009	0.955	0.118	0.403	0.978	0.075	0.052	0.948	0.125	0.008	NC	NC	NC
348	Hungary	0.956	0.108	0.010	0.926	0.135	0.407	0.989	0.048	0.044	0.966	0.082	0.012	0.854	0.081	0.413
352	Iceland	0.949	0.108	0.028	0.990	0.048	0.060	0.977	0.067	0.021	0.971	0.085	0.020	NC	NC	NC
360	Indonesia	0.946	0.063	0.023	0.958	0.070	0.052	0.935	0.058	0.092	0.790	0.151	0.114	0.871	0.054	0.165
372	Ireland	0.948	0.117	0.018	0.977	0.080	0.051	0.982	0.064	0.020	0.955	0.106	0.007	0.824	0.092	0.130
376	Israel	0.964	0.099	0.028	NA	NA	NA	0.952	0.100	0.077	0.941	0.110	0.009	0.908	0.072	0.150
380	Italy	0.976	0.064	0.010	NA	NA	NA	0.982	0.056	0.050	0.968	0.075	0.017	NC	NC	NC
392	Japan	0.973	0.071	0.019	0.983	0.047	0.073	0.976	0.074	0.072	0.828	0.184	0.100	NC	NC	NC
400	Jordan	0.956	0.077	0.052	0.990	0.053	0.011	0.986	0.051	0.027	0.968	0.070	0.052	0.918	0.057	0.111
410	Korea	0.974	0.072	0.030	0.891	0.178	0.123	0.982	0.061	0.023	0.928	0.132	0.033	0.873	0.094	0.145
411	Kosovo	0.962	0.070	0.035	0.962	0.118	0.011	NA	NA	NA	0.954	0.095	0.051	NC	NC	NC
422	Lebanon	0.974	0.040	0.032	0.996	0.023	0.028	0.957	0.053	0.071	0.962	0.062	0.034	0.906	0.040	0.190
428	Latvia	0.967	0.088	0.011	1.000	0.011	0.050	0.972	0.074	0.029	0.963	0.091	0.045	0.863	0.075	0.454
440	Lithuania	0.977	0.077	0.007	0.997	0.031	0.052	0.977	0.070	0.028	0.943	0.126	0.006	NC	NC	NC
442	Luxembourg	0.986	0.056	0.015	0.824	0.278	0.111	0.959	0.094	0.062	0.885	0.146	0.016	0.908	0.070	0.201
446	Macao	0.956	0.084	0.108	0.975	0.077	0.238	0.972	0.076	0.062	0.971	0.079	0.020	0.825	0.087	0.442
470	Malta	0.966	0.089	0.024	0.999	0.019	0.055	0.988	0.055	0.017	0.925	0.140	0.013	0.842	0.079	0.375
484	Mexico	0.979	0.054	0.034	0.886	0.179	0.117	0.986	0.050	0.021	0.976	0.080	0.007	0.863	0.084	0.417
498	Moldova	0.961	0.065	0.058	0.969	0.073	0.070	0.977	0.046	0.029	0.850	0.150	0.759	0.901	0.043	0.242

Table 7 (continued)

PISA code	Region/country	Classroom management			Teacher-directed instruction			Emotional support			Feedback			Inquiry-based instruction		
		CFI	RMSEA	SRMR between	CFI	RMSEA	SRMR between	CFI	RMSEA	SRMR between	CFI	RMSEA	SRMR between	CFI	RMSEA	SRMR between
499	Montenegro	0.975	0.075	0.034	0.984	0.075	0.006	0.979	0.077	0.004	0.943	0.120	0.020	0.929	0.073	0.029
528	Netherlands	0.985	0.052	0.025	0.982	0.051	0.155	0.984	0.054	0.051	0.976	0.084	0.013	0.899	0.079	0.400
554	New Zealand	0.953	0.123	0.010	0.982	0.081	0.085	0.930	0.143	0.057	0.963	0.101	0.011	0.859	0.093	0.422
578	Norway	0.970	0.082	0.023	0.986	0.063	0.108	0.985	0.060	0.008	0.946	0.131	0.005	0.901	0.079	0.261
604	Peru	0.954	0.081	0.066	0.781	0.255	0.285	0.993	0.035	0.022	0.958	0.086	0.038	0.900	0.068	0.258
616	Poland	0.965	0.088	0.008	0.959	0.122	0.230	0.977	0.076	0.025	0.624	0.307	0.040	0.892	0.078	0.188
620	Portugal	0.976	0.079	0.022	NA	NA	NA	0.984	0.063	0.017	0.954	0.117	0.018	NC	NC	NC
634	Qatar	0.951	0.088	0.018	0.991	0.041	0.050	0.974	0.061	0.065	0.949	0.098	0.039	0.905	0.068	0.203
642	Romania	0.956	0.048	0.025	0.936	0.060	0.098	0.977	0.038	0.123	NA	NA	NA	NC	NC	NC
643	Russian Federation	0.968	0.090	0.010	0.932	0.135	0.093	0.975	0.067	0.013	0.962	0.090	0.032	0.899	0.077	0.422
702	Singapore	0.942	0.120	0.027	0.996	0.039	0.055	0.961	0.099	0.034	0.956	0.110	0.016	0.835	0.098	0.141
703	Slovak Republic	0.950	0.099	0.023	0.931	0.133	0.186	0.984	0.054	0.025	0.886	0.159	0.025	0.889	0.078	0.189
704	Vietnam	0.952	0.047	0.051	0.951	0.073	0.125	0.976	0.037	0.054	0.868	0.114	0.079	0.853	0.062	0.287
705	Slovenia	0.974	0.076	0.012	NC	NC	NC	0.978	0.068	0.050	NC	NC	NC	NC	NC	NC
724	Spain	0.987	0.052	0.011	0.991	0.037	0.025	0.983	0.060	0.020	0.966	0.094	0.015	0.886	0.069	0.195
752	Sweden	0.964	0.084	0.027	0.979	0.086	0.044	0.989	0.054	0.010	0.951	0.126	0.011	0.833	0.109	0.140
756	Switzerland	0.989	0.049	0.035	0.901	0.175	0.217	0.985	0.052	0.051	0.963	0.087	0.018	NC	NC	NC
764	Thailand	0.959	0.074	0.028	0.977	0.090	0.048	0.953	0.095	0.081	0.955	0.085	0.030	0.910	0.074	0.077
780	Trinidad and Tobago	0.966	0.069	0.024	0.997	0.032	0.059	0.994	0.035	0.018	0.966	0.092	0.020	NC	NC	NC
784	United Arab Emirates	0.953	0.095	0.037	0.990	0.057	0.059	0.986	0.054	0.032	0.960	0.099	0.020	0.914	0.071	0.332

Table 7 (continued)

PISA code	Region/country	Classroom management			Teacher-directed instruction			Emotional support			Feedback			Inquiry-based instruction		
		CFI	RMSEA	SRMR between	CFI	RMSEA	SRMR between	CFI	RMSEA	SRMR between	CFI	RMSEA	SRMR between	CFI	RMSEA	SRMR between
788	Tunisia	0.964	0.067	0.024	0.970	0.093	0.050	0.978	0.065	0.016	0.925	0.124	0.045	0.906	0.067	0.212
792	Turkey	0.956	0.095	0.015	0.988	0.056	0.029	0.976	0.072	0.029	0.927	0.130	0.051	NC	NC	NC
807	FYROM	0.976	0.056	0.043	0.970	0.080	0.029	0.972	0.058	0.208	0.722	0.222	0.344	0.929	0.050	0.173
826	UK	0.955	0.111	0.020	0.989	0.058	0.420	0.961	0.100	0.031	0.975	0.079	0.015	0.881	0.080	0.469
840	USA	0.947	0.116	0.008	0.988	0.068	0.100	0.956	0.101	0.054	0.975	0.088	0.006	0.885	0.087	0.187
858	Uruguay	0.969	0.081	0.023	0.944	0.128	0.167	0.981	0.067	0.011	0.945	0.112	0.029	0.924	0.064	0.416
970	B-S-J-G (China)	0.959	0.086	0.026	0.931	0.142	0.084	0.975	0.060	0.016	0.941	0.118	0.011	0.911	0.080	0.099
971	Spain (regions)	0.985	0.053	0.015	0.993	0.032	0.036	0.986	0.057	0.008	0.969	0.085	0.005	0.875	0.074	0.138
974	Argentina (Ciudad Autónoma de Buenos)	0.994	0.034	0.025	0.992	0.036	0.173	0.967	0.076	0.049	0.954	0.104	0.023	0.856	0.070	0.302

NA, not applicable due to unavailable data; *NC*, data available but model estimation failed to converge

Table 8 Relations between teaching quality in PISA 2015 and students' intrinsic motivation to learn science

PISA code	Country/region	Adaptive instruction		Classroom management		Teacher-directed instruction		Emotional support		Feedback		Inquiry-based instruction	
		Beta (SE)	*p*	Beta (SE)	*p*	Beta (SE)	*p*	Beta (SE)	*p*	Beta (SE)	*p*	Beta (SE)	*p*
12	Algeria	NA	NA	0.57 (0.12)	0	0.57 (0.13)	0	NC	NC	− 0.64 (0.28)	0.122	− 0.57 (0.17)	0.72
36	Australia	0.76 (0.06)	0	0.64 (0.05)	0	0.79 (0.08)	0	0.64 (0.07)	0	0.07 (0.11)	0.557	NC	NC
40	Austria	0.37 (0.14)	0.002	0.47 (0.08)	0	0.92 (0.08)	0	0.37 (0.17)	0.054	− 0.34 (0.10)	0.027	0.12 (0.12)	0.282
56	Belgium	1.21 (0.27)	0.122	0.17 (0.11)	0.114	1.02 (0.05)	0	0.59 (0.08)	0.003	0.32 (0.11)	0.003	0.66 (0.11)	0.001
76	Brazil	0.82 (0.06)	0	0.66 (0.05)	0	0.82 (0.07)	0	0.58 (0.12)	0	0.20 (0.13)	0.218	− 0.28 (0.08)	0.021
100	Bulgaria	0.77 (0.10)	0	0.82 (0.06)	0	1.00 (0.22)	0.017	− 0.38 (0.22)	0.076	− 0.81 (0.13)	0.039	− 0.81 (0.10)	0
124	Canada	0.57 (0.10)	0	− 0.02 (0.11)	0.866	0.74 (0.07)	0	0.33 (0.16)	0.154	NC	NC	NC	NC
152	Chile	0.71 (0.12)	0	0.60 (0.09)	0	0.79 (0.08)	0	0.22 (0.15)	0.181	− 0.29 (0.15)	0.052	− 0.47 (0.22)	0.384
158	Chinese Taipei	0.32 (0.19)	0.151	0.40 (0.10)	0	0.73(0.08)	0	0.58 (0.11)	0.029	− 0.42 (0.16)	0.026	NC	NC
170	Colombia	0.15 (0.26)	0.631	0.47 (0.11)	0.007	− 0.60 (0.10)	0.002	0.52 (0.10)	0.002	0.59 (0.10)	0.014	NC	NC
188	Costa Rica	0.78 (0.19)	0	0.37 (0.21)	0.046	0.67 (0.33)	0.75	0.78 (0.16)	0	1 (0.15)	0	0.73 (0.37)	0.495
191	Croatia	0.31 (0.18)	0.063	0.72 (0.07)	0	0.69 (0.11)	0	0.20 (0.18)	0.287	−0.45 (0.10)	0	−0.04 (0.24)	0.859
203	Czech Republic	0.33 (0.12)	0.003	0.73 (0.05)	0	0.80 (0.15)	0	− 0.05 (0.16)	0.749	NC	NC	NC	NC
208	Denmark	0.69 (0.23)	0	0.52 (0.24)	0.017	− 0.17 (0.20)	0.369	0.44 (0.85)	0.754	0.35 (0.14)	0.099	0.53 (0.19)	0.044
214	Dominican Republic	0.96 (0.39)	0.002	0.14 (0.93)	0.955	NC	NC	0.55 (0.29)	0.805	0.83 (0.32)	0.58	NC	NC
233	Estonia	0.27 (0.15)	0.099	0.47 (0.13)	0.001	0.78 (0.20)	0.021	0.72 (0.11)	0	− 0.69 (0.22)	0.003	NC	NC
246	Finland	0.78 (0.10)	0	0.75 (0.08)	0	NC	NC	0.78 (0.09)	0	0.26 (0.23)	0.29	NC	NC
250	France	− 0.19 (0.25)	0.53	0.60 (0.09)	0	0.89 (0.09)	0	− 0.24 (0.16)	0.117	− 0.61 (0.11)	0	NC	NC
268	Georgia	NA	NA	0.69 (0.09)	0	0.65 (0.22)	0.107	0.50 (0.14)	0.006	NC	NC	0.24 (0.30)	0.25
276	Germany	0.97 (0.10)	0	0.84 (0.07)	0	0.99 (0.09)	0	− 0.65 (0.23)	0.301	− 0.76 (0.12)	0.217	0.80 (0.10)	0.001
300	Greece	0.33 (0.17)	0.029	0.84 (0.06)	0	0.85 (0.10)	0	0.03 (0.25)	0.912	− 0.62 (0.12)	0	− 0.33 (0.16)	0.239

Table 8 (continued)

PISA code	Country/region	Adaptive instruction		Classroom management		Teacher-directed instruction		Emotional support		Feedback		Inquiry-based instruction	
		Beta (SE)	*p*	Beta (SE)	*p*	Beta (SE)	*p*	Beta (SE)	*p*	Beta (SE)	*p*	Beta (SE)	*p*
344	Hong Kong	NA	NA	0.37 (0.13)	0.077	0.71 (0.12)	0.004	NC	NC	NC	NC	NC	NC
348	Hungary	0.46 (0.18)	0.023	0.90 (0.06)	0	0.77 (0.10)	0	− 0.10 (0.35)	0.773	− 0.65 (0.12)	0	− 0.13 (0.18)	0.539
352	Iceland	0.37 (0.17)	0.026	0.29 (0.18)	0.113	0.68 (0.16)	0.005	0.31 (0.19)	0.096	− 0.27 (0.19)	0.164	NC	NC
360	Indonesia	NA	NA	0.44 (0.09)	0	0.28 (0.14)	0.068	0.55 (0.15)	0.002	0.13 (0.16)	0.416	0.09 (0.17)	0.566
372	Ireland	0.89 (0.18)	0	0.42 (0.14)	0.002	0.79 (0.13)	0	0.90 (0.11)	0	0.30 (0.50)	0.532	0.68 (0.20)	0.011
376	Israel	0.80 (0.08)	0	0.63 (0.10)	0	NC	NC	0.54 (0.10)	0	0.52 (0.09)	0	0.69 (0.07)	0
380	Italy	0.41 (0.09)	0	0.51 (0.05)	0	NC	NC	0.17 (0.07)	0.014	− 0.02 (0.06)	0.793	NC	NC
392	Japan	0.44 (0.08)	0	0.43 (0.08)	0	0.71 (0.08)	0	− 0.15 (0.14)	0.528	− 0.17 (0.10)	0.289	NC	NC
400	Jordan	NA	NA	0.69 (0.07)	0	0.91 (0.03)	0	0.67 (0.21)	0.254	1.03 (0.09)	0	− 0.29 (0.13)	0.074
410	Korea	0.40 (0.14)	0.003	0.54 (0.10)	0	0.72 (0.13)	0.419	0.51 (0.11)	0	− 0.41 (0.22)	0.046	− 0.19 (0.18)	0.369
411	Kosovo	NA	NA	0.63 (0.11)	0	0.66 (0.11)	0	NC	NC	0.23 (0.19)	0.251	NC	NC
422	Lebanon	NA	NA	0.81 (0.06)	0	0.86 (0.05)	0	0.86 (0.09)	0.344	0.49 (0.09)	0	− 0.57 (0.14)	0.436
428	Latvia	1.02 (0.19)	0.82	NC	NC	NC	NC	0.58 (0.18)	0.023	NC	NC	− 1.05 (0.12)	0.246
440	Lithuania	− 0.86 (0.12)	0.013	0.60 (0.08)	0	NC	NC	NC	NC	− 0.75 (0.11)	0	NC	NC
442	Luxembourg	0.07 (0.43)	0.862	0.63 (0.16)	0	0.74 (0.09)	0.001	0.01 (0.33)	0.973	− 0.40 (0.26)	0.134	− 0.39 (0.36)	0.493
446	Macao	0.56 (0.39)	0.343	0.74 (0.22)	0	NC	NC	0.52 (0.19)	0.018	− 0.36 (0.32)	0.293	− 0.40 (0.48)	0.514
470	Malta	NA	NA	0.80 (0.07)	0	0.84 (0.07)	0	0.44 (0.17)	0.008	− 0.11 (0.13)	0.44	0.87 (0.14)	0.063
484	Mexico	NA	NA	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
498	Moldova	NA	NA	NC	NC	NC	NC	NC	NC	0.58 (0.12)	0.007	NC	NC
499	Montenegro	0.45 (0.19)	0.046	0.86 (0.09)	0	0.58 (0.15)	0	0.28 (0.18)	0.128	0.07 (0.43)	0.856	0.04 (0.27)	0.867
528	Netherlands	0.63 (0.13)	0.126	0.52 (0.09)	0	0.96 (0.07)	0	− 0.64 (0.12)	0	− 0.65 (0.12)	0	− 0.77 (0.08)	0.029

Table 8 (continued)

PISA code	Country/region	Adaptive instruction		Classroom management		Teacher-directed instruction		Emotional support		Feedback		Inquiry-based instruction	
		Beta (SE)	*p*	Beta (SE)	*p*	Beta (SE)	*p*	Beta (SE)	*p*	Beta (SE)	*p*	Beta (SE)	*p*
554	New Zealand	0.73 (0.14)	0.001	0.70 (0.12)	0	0.66 (0.14)	0	−0.51 (0.21)	0.189	−0.25 (0.42)	0.465	−0.88 (0.11)	0.792
578	Norway	0.79 (0.10)	0	0.53 (0.12)	0	0.67 (0.14)	0	0.61 (12)	0	0.56 (0.17)	0.002	−0.02 (0.25)	0.949
604	Peru	NA	NA	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
616	Poland	0.41 (0.23)	0.086	0.64 (0.13)	0	NC	NC	0.35 (0.25)	0.196	−0.86 (0.42)	0.678	−0.18 (0.40)	0.705
620	Portugal	−0.19 (0.32)	0.54	0.28 (0.17)	0.368	NC	NC	−0.69 (0.13)	0.429	−0.93 (0.09)	0.509	NC	NC
634	Qatar	0.81 (0.07)	0	0.73 (0.06)	0	0.81 (0.07)	0	0.62 (0.14)	0.053	0.58 (0.31)	0.013	−0.48 (0.09)	0.008
642	Romania	NA	NA	0.76 (0.08)	0	0.86 (0.07)	0	0.32 (0.36)	0.727	NC	NC	NC	NC
643	Russian Federation	0.74 (0.13)	0	0.68 (0.10)	0	NC	NC	0.40 (0.24)	0.049	0.14 (0.30)	0.656	−0.86 (0.19)	0.613
702	Singapore	0.85 (0.07)	0	0.69 (0.10)	0	0.96 (0.06)	0	0.72 (0.10)	0	−0.14 (0.16)	0.382	0.65 (0.10)	0.006
703	Slovak Republic	0.97 (0.54)	0.46	0.72 (0.06)	0	0.12 (0.13)	0.318	−0.20 (0.14)	0.178	−0.44 (0.10)	0	−0.15 (0.25)	0.546
704	Vietnam	NA	NA	−0.06 (0.43)	0.888	0.58 (0.38)	0.086	0.30 (0.27)	0.355	NC	NC	−0.53 (0.22)	0.062
705	Slovenia	NA	NA	0.63 (0.08)	0	NA	NA	0.07 (1.01)	0.946	NA	NA	NC	NC
724	Spain	0.29 (0.18)	0.128	0.36 (0.13)	0.008	0.55 (0.15)	0.001	0.32 (0.14)	0.04	−0.05 (0.14)	0.721	−0.12 (0.22)	0.636
752	Sweden	0.49 (0.11)	0	0.61 (0.10)	0	0.64 (0.11)	0	0.27 (0.12)	0.053	−0.07 (0.66)	0.916	0.12 (0.22)	0.665
756	Switzerland	0.27 (0.12)	0.046	0.33 (0.13)	0.025	0.89 (0.05)	0	0.37 (11)	0.006	−0.36 (0.15)	0.028	NC	NC
764	Thailand	0.64 (0.13)	0.161	0.70 (0.09)	0	0.87 (0.10)	0	0.57 (0.13)	0	0.25 (0.22)	0.298	0.35 (0.17)	0.027
780	Trinidad and Tobago	NA	NA	0.80 (0.08)	0	0.67 (0.09)	0	0.45 (0.14)	0.009	−0.37 (0.21)	0.103	NC	NC
784	United Arab Emirates	0.73 (0.06)	0	0.64 (0.07)	0	0.84 (0.21)	0.002	0.42 (0.09)	0	0.49 (0.11)	0	−0.30 (0.14)	0.174
788	Tunisia	0.58 (0.22)	0.141	0.51 (0.15)	0.045	0.79 (0.16)	0	−0.03 (0.18)	0.884	NC	NC	−0.74 (0.14)	0.651
792	Turkey	0.96 (0.06)	0	0.71 (0.09)	0	0.80 (0.12)	0	0.05 (0.42)	0.911	−0.27 (0.19)	0.662	NC	NC
807	FYROM	NA	NA	0.58 (0.13)	0	0.58 (0.16)	0.004	NC	NC	0.13 (0.23)	0.535	−0.13 (0.19)	0.605

Table 8 (continued)

PISA code	Country/region	Adaptive instruction		Classroom management		Teacher-directed instruction		Emotional support		Feedback		Inquiry-based instruction	
		Beta (SE)	*p*	Beta (SE)	*p*	Beta (SE)	*p*	Beta (SE)	*p*	Beta (SE)	*p*	Beta (SE)	*p*
826	UK	0.67 (0.14)	0.11	0.51 (0.71)	0.779	0.41 (0.28)	0.231	− 0.28 (0.16)	0.147	− 0.17 (0.11)	0.444	0.01 (0.11)	0.947
840	USA	0.76 (0.18)	0.004	0.69 (0.12)	0	0.60 (0.24)	0.143	0.72 (0.24)	0.007	0.10 (0.27)	0.728	0.11 (0.31)	0.581
858	Uruguay	0.15 (0.61)	0.815	0.65 (0.15)	0.001	0.93 (0.04)	0	− 0.16 (0.34)	0.641	− 0.43 (0.19)	0.171	− 0.44 (0.20)	0.481
970	B-S-J-G (China)	0.89 (0.04)	0	0.83 (0.04)	0	0.33 (0.17)	0.014	0.81 (0.05)	0	0.50 (0.09)	0	0.70 (0.06)	0
971	Spain (regions)	0.36 (0.06)	0	0.40 (0.05)	0	0.38 (0.46)	0.965	0.32 (0.06)	0	0.02 (0.08)	0.853	− 0.04 (0.10)	0.729
974	Argentina (Ciudad Autónoma de Buenos)	NA	NA	0.56 (0.26)	0.07	NC	NC	0.67 (0.36)	0.553	0.80 (0.13)	0.283	0.11 (0.40)	0.858

NA, not applicable due to unavailable data; *NC*, data available but model estimation failed to converge

References

Aditomo, A. (2020). Science teaching practices in Indonesian secondary schools: a portrait of educational quality and equity based on PISA 2015. In A. Suryani, H. Masalam, & I. Tirtowalujo (Eds.), *Preparing Indonesian youth: a review of educational research*. Brill: Leiden.

Aditomo, A., & Klieme, E. (2020). Forms of inquiry-based science instruction and their relations with learning outcomes: evidence from high and low-performing education systems. *International Journal of Science Education, 42*(4), 504–525. https://doi.org/10.1080/09500693.2020.1716093.

Aldrup, K., Klusmann, U., Lüdtke, O., Göllner, R., Trautwein, U., Aldrup, K., & Klusmann, U. (2018). Social support and classroom management are related to secondary students' general school adjustment: a multilevel structural equation model using student and teacher ratings. *Journal of Educational Psychology., 110*, 1066–1083.

Blazar, D. (2015). Effective teaching in elementary mathematics: identifying classroom practices that support student achievement. *Economics of Education Review, 48*, 16–29. https://doi.org/10.1016/j.econedurev.2015.05.005.

Blazar, D., & Kraft, M. A. (2017). Teacher and teaching effects on students' attitudes and behaviors. *Educational Evaluation and Policy Analysis, 39*(1), 146–170. https://doi.org/10.3102/0162373716670260.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How people learn: brain, mind, experience, and school*. Washington DC: National Academies Press.

Breakspear, S. (2012). The policy impact of PISA: an exploration of the normative effects of international benchmarking in school system performance. OECD Education Working Papers, No. 71. *OECD Publishing (NJ1)*, (71). Retrieved from http://eric.ed.gov/?id=ED530643.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research second edition* (2nd ed.). New York: The Guilford Press.

Carroll, J. B. (1989). The Carroll Model: a 25-year retrospective and prospective view. *Educational Researcher, 18*(1), 26–31.

Cooley, W., & Leinhardt, G. (1980). The instructional dimensions study. *Educational Evaluation and Policy Analysis, 2*, 7–25.

Creemers, B. P. M. (1994). *The effective classroom*. New York: Cassell.

Deci, E. L., Ryan, R. M., Vallerand, R. J., & Pelletier, L. G. (1991). Motivation and education: the self-determination perspective. *Educational Psychologist, 26*(3–4), 325–346. https://doi.org/10.1080/00461520.1991.9653137.

Deci, E. L., Ryan, R. M., & Williams, G. C. (1996). Need satisfaction and the self-regulation of learning. *Learning and Individual Differences, 8*(3), 165–183. https://doi.org/10.1016/S1041-6080(96)90013-8.

Decristan, J., Kunter, M., Fauth, B., Büttner, G., Hardy, I., & Hertel, S. (2016). What role does instructional quality play for elementary school children's science competence?: a focus on students at risk/Zur Bedeutung von Unterrichtsqualität für die naturwissenschaftliche Kompetenz von Grundschulkindern. *Journal for Educational Research Online, 8*(1), 66–89.

Derry, S. J. (1996). Cognitive schema theory in the constructivist debate. *Educational Psychologist, 31*(3–4), 163–174. https://doi.org/10.1080/00461520.1996.9653264.

Egeberg, H. M., Mcconney, A., & Price, A. (2016). Classroom management and national professional standards for teachers: a review of the literature on theory and practice. *Australian Journal of Teacher Education, 41*(7), 1–18.

Emmer, E. T., & Stough, L. M. (2001). Classroom management: a critical part of educational psychology, with implications for teacher education. *Educational Psychologist, 36*(2), 103–112. https://doi.org/10.1207/S15326985EP3602.

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.

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.

Gee, K. A., & Wong, K. K. (2012). A cross national examination of inquiry and its relationship to student performance in science: evidence from the Program for International Student Assessment (PISA) 2006. *International Journal of Educational Research, 53*, 303–318. https://doi.org/10.1016/j.ijer.2012.04.004.

Gershenson, S. (2016). Linking teacher quality, student attendance, and student achievement. *Education Finance and Policy, 11*, 125–149.

Goldstein, H. (1997). Methods in school effectiveness research. *School Effectiveness and School Improvement, 8*, 369–395.

Grek, S. (2009). Governing by numbers: The PISA "effect" in Europe. *Journal of Education Policy, 24*(1), 23–37. https://doi.org/10.1080/02680930802412669.

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* (pp. 39–56). Charlotte, NC: Information Age Publishing.

Hox, J. J. (2010). 从汶川地震到芦山地震. In *Multilevel analysis: techniques and applications* (2nd ed.). New York: Routledge. https://doi.org/10.1360/zd-2013-43-6-1064.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal1, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118.

Hutchinson, D., & Healy, M. (2001). The effect on variance component estimates of ignoring a level in a multilevel model. *Multilevel Modeling Newsletter, 13*(2), 4–5.

Hwang, J., Choi, K. M., Bae, Y., & Shin, D. H. (2018). Do teachers' instructional practices moderate equity in mathematical and scientific literacy?: an investigation of the PISA 2012 and 2015. *International Journal of Science and Mathematics Education, 16*, 25–45. https://doi.org/10.1007/s10763-018-9909-8.

Jackson C. K. (2012). Non-cognitive ability, test scores, and teacher quality: evidence from 9th grade teachers in North Carolina (NBER Working Paper No. 18624). Cambridge, MA: National Bureau for Economic Research.

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Estimating Within-Group Interrater Reliability with and without Response Bias, 69*(1), 85–98.

Jennings, J. L., & Di Prete, T. A. (2010). Teacher effects on social and behavioral skills in early elementary school. *Sociology of Education, 83*, 135–159.

Jiang, F., & McComas, W. F. (2015). The effects of inquiry teaching on student science achievement and attitudes: evidence from propensity score analysis of PISA data. *International Journal of Science Education, 37*(3), 554–576. https://doi.org/10.1080/09500693.2014.1000426.

Kelloway, E. K. (2015). *Using Mplus for structural equation modeling* (2nd ed.). Thousand Oaks: Sage Publications.

Klein, K. J., Bliese, P. D., Kozlowski, S. W. J., Dansereau, F., Gavin, M. B., Griffin, M. A., … Bligh, M. C. (2000). Multilevel analytical techniques: commonalities, differences, and continuing questions. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations, and methods in organizations: foundations, extensions, and new directions* (pp. 512–553). San Francisco: Jossey-Bass.

Klieme, E. (2013). The role of large-scale assessments in research on educational effectiveness and school development. In M. von Davier, E. Gonzalez, I. Kirsch, & K. Yamamoto (Eds.), *The role of international large-scale assessments: perspectives from technology, economy, and educational research* (pp. 115–148). Dordrecht: Springer.

Klieme, E., Pauli, C., & Reusser, K. (2009). *The Pythagoras study: investigating effects of teaching and learning in Swiss and German mathematics classrooms* (pp. 137–160). Munster: Waxmann Verlag.

Kraft, M. A., & Grace, S. (2016). *Teaching for tomorrow's economy? Teacher effects on complex cognitive skills and social-emotional competencies (working paper)*. Providence, RI: Brown University Retrieved from http://scholar.harvard.edu/files/mkraft/files/teaching_for_tomorrows_economy_-_final_public.pdf.

Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research, 9*(3), 231–251. https://doi.org/10.1007/s10984-006-9015-7.

Kyriakides, L., Campbell, R. J., & Gagatsis, A. (2000). The significance of the classroom effect in primary schools: an application of Creemers' comprehensive model of educational effectiveness. *School Effectiveness and School Improvement, 11*, 501–529.

Lau, K., & Lam, T. Y. (2017). Instructional practices and science performance of 10 top-performing regions in PISA 2015. *International Journal of Science Education, 39*, 1–22. https://doi.org/10.1080/09500693.2017.1387947.

Lazonder, A. W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning: effects of guidance. *Review of Educational Research, 86*(3), 681–718. https://doi.org/10.3102/0034654315627366.

Lebreton, J. M., & Senter, J. L. (2008). Answers to 20 questions and interrater agreement. *Organizational Research Methods, 11*(4), 815–852. https://doi.org/10.1177/1094428106296642.

Lüdtke, O., Trautwein, U., Kunter, M., & Baumert, J. (2006). Reliability and agreement of student ratings of the classroom environment: a reanalysis of TIMSS data. *Learning Environments Research, 9*(3), 215–230. https://doi.org/10.1007/s10984-006-9014-8.

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*(2), 120–131. https://doi.org/10.1016/j.cedpsych.2008.12.001.

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2 x 2 taxonomy of multilevel latent contextual models: accuracy-bias trade-offs in full and partial error correction models. *Psychological Methods, 16*(4), 444–467. https://doi.org/10.1037/a0024376.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: the critical issues of validity, bias, and utility. *American Psychologist, 52*(11), 1187–1197.

Marsh, H. W., Hau, K-T., & Wen, Z. (2004). In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler's (1999) Findings. *Structural Equation Modeling: A Multidisciplinary Journal, 11*(3), 320–341. https://doi.org/10.1207/s15328007sem1103_2.

Martínez, J.F. (2012). Consequences of omitting the classroom in multilevel models of schooling: an illustration using opportunity to learn and reading achievement. *School Effectiveness and School Improvement, 23*(13), 305–326.

McConney, A., Oliver, M. C., Conney, A. W.-M., Schibeci, R., & Maor, D. (2014). Inquiry, engagement, and literacy in science: a retrospective, cross-national analysis using PISA 2006. *Science Education, 98*, 963–980. https://doi.org/10.1002/sce.21135.

Miller, A. D., & Murdock, T. B. (2007). Modeling latent true scores to determine the utility of aggregate student perceptions as classroom indicators in HLM: the case of classroom goal structures. *Contemporary Educational Psychology, 32*, 83–104. https://doi.org/10.1016/j.cedpsych.2006.10.006.

Moerbeek, M. (2004). The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research, 39*, 127–149.

Morin, A. J. S., Marsh, H. W., Nagengast, B., & Scalas, L. F. (2013). Doubly latent multilevel analyses of classroom climate: an illustration. *Journal of Experimental Education., 82*, 143–167. https://doi.org/10.1080/00220973.2013.769412.

Müller, K., Prenzel, M., Seidel, T., Schiepe-tiska, A., & Kjærnsli, M. (2016). Science teaching and learning in schools: theoretical and empirical foundations for investigating classroom-level processes. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Assessing contexts of learning: an international perspective* (pp. 423–446). Switzerland: Springer International. https://doi.org/10.1007/978-3-319-45357-6.

Muthe'n, B. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28*, 338–354.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Los Angeles: Muthén & Muthén.

OECD. (2016). *PISA 2015 assessment and analytical framework: science, reading, mathematic and financial literacy*. Paris: OECD Publishing. https://doi.org/10.1787/9789264255425-en.

OECD. (2017). *PISA 2015 technical report*. Paris: OECD Publishing.

Perry, J. L., Nicholls, A. R., Clough, P. J., & Crust, L. (2015). Assessing model fit: caveats and recommendations for confirmatory factor analysis and exploratory structural equation modeling. *Measurement in Physical Education and Exercise Science, 19*(1), 12–21, https://doi.org/10.1080/1091367X.2014.952370.

Praetorius, A. K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: the German framework of three basic dimensions. *ZDM, 50*(3), 407–426. https://doi.org/10.1007/s11858-018-0918-4.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: applications and data analysis methods*. Thousand Oaks: Sage.

Raudensbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: applications and data analysis methods* (2nd ed.). Thousand Oaks: Sage.

Rjosk, C., Richter, D., Hochweber, J., Lüdtke, O., Klieme, E., & Stanat, P. (2014). Socioeconomic and language minority classroom composition and individual reading achievement: the mediating role of instructional quality. *Learning and Instruction, 32*, 63–72. https://doi.org/10.1016/j.learninstruc.2014.01.007.

Rudasill, K. M., Snyder, K. E., Levinson, H., & Adelson, J. L. (2018). Systems view of school climate: a theoretical framework for research. *Educational Psychology Review, 30*(35), 35–60. https://doi.org/10.1007/s10648-017-9401-y.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68.

Scheerens, J. & Creemers, BPM. (1989). Conceptualizing school effectiveness. *International Journal of Educational Research, 13*(7), 691–706.

Schiepe-tiska, A. (2019). School tracks as differential learning environments moderate the relationship between teaching quality and multidimensional learning goals in mathematics. *Frontiers in Education, 4*(January). https://doi.org/10.3389/feduc.2019.00004.

Schweig, J. (2014). Cross-level measurement invariance in school and classroom environment surveys: implications for policy and practice. *Educational Evaluation and Policy Analysis, 36*(3), 259–280. https://doi.org/10.3102/0162373713509880.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428.

Slater, H., Davies, N. M., & Burgess, S. (2012). Do Teachers Matter? Measuring the Variation in Teacher Effectiveness in England. *Oxford Bulletin of Economics and Statistics, 75*(4), 629–645. https://doi.org/10.1111/j.1468-0084.2011.00666.x.

Stapleton, L. M., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics, 41*(5), 481–520. https://doi.org/10.3102/1076998616646200.

Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics, 41*(5), 481–520. https://doi.org/10.3102/1076998616646200.

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

Wagner, W., Göllner, R., Werth, S., Voss, T., Schmitz, B., & Trautwein, U. (2016). Student and teacher ratings of instructional quality: consistency of ratings over time, agreement, and predictive power. *Journal of Educational Psychology, 108*(5), 705–721.

Wallace, T. L., Kelcey, B., & Ruzek, E. (2016). What can student perception surveys tell us about teaching? Empirically testing the underlying structure of the tripod student perception survey. *American Educational Research Journal, XX*(X), 1–35. https://doi.org/10.3102/0002831216671864.

Wang, M., & Degol, J. L. (2016). School climate: a review of the construct, measurement, and impact on student outcomes. *Educational Psychology Review, 38*, 315–352. https://doi.org/10.1007/s10648-015-9319-1.

Wenger, M., Lüdtke, O., & Brunner, M. (2018). Übereinstimmung, Variabilität und Reliabilität von Schülerurteilen zur Unterrichtsqualität auf Schulebene: Ergebnisse aus 81 Ländern. *Zeitschrift Fur Erziehungswiss, 21*, 929–950. https://doi.org/10.1007/s11618-018-0813-3.

Wittrock, M. C. (2010). Learning as a generative process. *Educational Psychologist, 45*(1), 40–45. https://doi.org/10.1080/00461520903433554.

Zyphur, M. J., Kaplan, S. A., & Christian, M. S. (2008). Assumptions of cross-level measurement and structural invariance in the analysis of multilevel data: problems and solutions. *Group Dynamics: Theory, Research, and Practice, 12*(2), 127–140. https://doi.org/10.1037/1089-2699.12.2.127.