Instructional quality: catalyst or pitfall in educational systems' aim for high achievement and equity? An answer based on multilevel SEM analyses of TIMSS 2015 data in Flanders (Belgium), Germany, and Norway

Kim Bellens[1]* , Jan Van Damme[1], Wim Van Den Noortgate[2], Heike Wendt[3] and Trude Nilsen[4]

*Correspondence:
kim.bellens@kuleuven.be
[1] Centre for Educational
Effectiveness and Evaluation,
Faculty of Psychology
and Educational Sciences,
KU Leuven, Dekenstraat
2, Box 3773, 3000 Leuven,
Belgium
Full list of author information
is available at the end of the
article

Abstract

In the Trends in Mathematics and Science Study (TIMSS) 2015, the educational systems from the Dutch speaking part of Belgium (Flanders), Germany, and Norway included scales capturing three dimensions of instructional quality (INQUA): classroom management, supportive climate, and cognitive activation. With the inclusion of these extra scales, a unique opportunity was created to investigate the various dimensions of INQUA and their relation to educational outcomes. In this study, multilevel structural equational modelling analyses are conducted to answer three research questions: (a) 'Do the items reliably measure the three dimensions of INQUA as classroom constructs? And if not, can we build reliable scales with the items, capturing the dimensions of INQUA as classroom constructs?' (b) 'To what extent can INQUA contribute to achievement?', and (c) 'To what extent can INQUA contribute to (social and ethnic) equity?'. Results indicate that INQUA might serve as a catalyst in increasing achievement in education systems. Furthermore, results indicate that INQUA does not relate to equity and consequently that all students benefit from the same educational practices. However, issues in the cross-cultural conceptualization and measurement of INQUA are raised, questioning the extent to which the three well-known dimensions of INQUA (a) are well-defined and might be sufficiently differentiated from each other, (b) sufficiently capture the diverse set of educational practices relating to students' educational outcomes, and (c) can be established across countries in a unified manner. The results of this study give direction to educational practitioners and policy makers in creating and providing effective learning environments.

Keywords: Instructional quality (INQUA), Outcomes of educational systems, Achievement and equity, TIMSS 2015, Multilevel SEM

Bellens *et al. Large-scale Assess Educ* (2019) 7:1

Page 2 of 27

Introduction

Worldwide, millions of students receive instruction in classrooms, day in day out. We assume (and demand) that the instructional practices that teachers set up have a beneficial impact on a diverse set of outcomes, ranging from individual benefits of schooling to broader societal goals like a competitive economy and societal equity. However, to date, a lot may still be discovered with regard to cross-cultural evidence in the practices underlying instructional quality (INQUA) and how those practices relate to educational outcomes. With this study, we aim to contribute to the current knowledge on instructional quality and provide direction to educational practitioners and policy-makers in creating and providing effective learning environments.

Background

INQUA: three dimensions

INQUA is 'a construct that reflects those features of teachers' instructional practices well known to be positively related to student outcomes, both cognitive and affective ones' (Nilsen and Gustafsson 2016, p. 5). INQUA is assumed to be an individual teacher's characteristic, going beyond only students' (biased) perception or evaluation of a single teaching performance (Berliner 2005). With INQUA being a key determinant of educational outcomes, the international educational research community has a large interest in studying this concept and its relation to a wide variety of educational outcomes (see e.g., Nilsen and Gustafsson 2016).

In the multitude of instructional practices included in INQUA, an agreement exists on three main dimensions: (a) classroom management (CM), (b) supportive climate (SC), and (c) cognitive activation (CA) (Fauth et al. 2014; Klieme et al. 2009; Kunter and Baumert 2006), referring to organizational, emotional, and instructional support, respectively (Pianta and Hamre 2009). The first dimension, CM, 'focuses on classroom rules and procedures, coping with disruptions, and smooth transitions' (Fauth et al. 2014, p. 2). This dimension focuses on giving opportunities to learn by making time and space for learning to take place, and in this sense, serves as a pre-condition for learning (Kunter and Baumert 2006; Klieme et al. 2009). The second dimension, SC, relates to 'supportive teacher–student relationships, positive and constructive teacher feedback, a positive approach to student errors and misconceptions, individual learner support and caring teacher behavior' (Klieme et al. 2009, p. 141). A third and final dimension is CA, described as the 'pedagogical practices used by teachers to promote student engagement in higher-level thinking' (Praetorius et al. 2014, p. 3), and focuses on enhancement of students' cognitive engagement which leads to learning (Fauth et al. 2014). It concerns the 'types of problems selected and the way in which they are implemented' (Kunter and Baumert 2006, p. 235).

Measuring INQUA: overview, pitfalls and possibilities

Despite the agreement of authors on the three theoretical dimensions of INQUA, the field of measuring INQUA is fragmented. In measuring INQUA, three main methods of data collection that each have their own pros and cons can be distinguished: teacher

Bellens *et al. Large-scale Assess Educ* (2019) 7:1

Page 3 of 27

self-ratings (see e.g., Kunter and Baumert 2006), observations (see e.g., Klieme et al. 2009, who used observations among other methods in the Pythagoras study), and student ratings.

One of the most well-known scales measuring all three constructs by means of student questionnaires in primary education is the scale from Fauth et al. (2014). Fauth et al. (2014) showed that student ratings are a reliable source in capturing teachers' INQUA in primary schools. This scale, however, has only been tested in Germany. Other existing INQUA scales have only been tested in samples of secondary school students (e.g., Kunter and Baumert 2006; PISA used various scales on the three dimensions of INQUA, OECD 2005, 2013) or are limited in the sense that they do not discriminate between the three different dimensions of INQUA (e.g., the scale 'Students' view on engaging teaching in mathematics lessons' captures some aspects of INQUA in the Trends in Mathematics and Science Study (TIMSS); Mullis et al. 2016). As noted limitations, teacher popularity might contaminate ratings students give of teachers (Fauth et al. 2014), and teacher ratings might also be contaminated by teaching ideals or self-serving strategies (Wubbels et al. 1992). Similarly, the presence of an external observer (e.g., researcher) might contaminate teachers' behaviors (Praetorius et al. 2012).

Comparing different existing scales of INQUA reveals a consensus on the content of the variables measuring CM (Praetorius et al. 2014). The operationalization of SC and CA, however, shows dispersion across those scales. As summarized by Praetorius et al. (2014), differences in the operationalization of SC measurement are depicted as only including climate variables (i.e., student–teacher relationship) or including a combination of aspects of climate and content-related support activities. Differences in the operationalization of CA measurement through various means can be mostly narrowed down to which course- and content-specificity is taken into account (Praetorius et al. 2014).

A lot of effort has been spent to establish measures of INQUA across countries in secondary education. For instance, based on PISA data, Scherer et al. (2016) confirmed the three factor structure of INQUA in three countries. To date, cross-cultural validation and data measuring the three dimensions of INQUA in primary education, however, is absent (O'Dwyer et al. 2015). Cross-cultural validation might give an additional benchmark to educational systems to evaluate the educational practices in their country in relation to the outcomes they are aiming for.

In light of this absence, a limited set of educational systems, i.e., the systems from the Dutch speaking part of Belgium (Flanders), Germany, and Norway, included existing scales in the student questionnaire in TIMSS 2015, capturing the three dimensions of INQUA. With the class-based sampling of TIMSS, this large-scale international assessment is optimally suited to investigate class and teacher practices across educational systems. Three main criteria were considered in the choice of items:[1] (1) maintain same items for assessment in both Grade 4 and Grade 8, (2) preference for items that were internationally tested and shown to be reliable and valid across countries, and (3) practical considerations of length to answer the (extended) questionnaire in a feasible

---

[1] The authors were involved in the selection of the scales and items.

timeframe. CM was measured by the PISA 2003 scale on CM (OECD 2005), of which the items can be found in Appendix. SC and CA items were taken from Fauth et al. (2014). For feasibility reasons, the SC scale was reduced from nine items to the five items that showed the highest factor loadings in Fauth et al. (2014). In Norway, the people in charge of the TIMSS 2015 data collection opted for different items on SC, stemming from the PISA 2012 questionnaire (OECD 2013) and Ferguson (2012), due to their practical consideration that these items measuring SC are more suitable for Grade 8 than the items of Fauth et al. (2014). In contrast to the SC scale of Fauth et al. (2014) which includes statements on the teacher as a person, mainly focusing on the student–teacher relationship, the SC scale in Norway mainly focuses on content-related support activities. An overview of the items included in the TIMSS 2015 student questionnaire of the three countries can be found in Appendix.

INQUA: contribution to educational systems' outcomes

It is assumed that educational outcomes are fostered by the effective interplay of the three dimensions of INQUA (Klieme et al. 2009). Two educational outcomes high on the political and societal agenda worldwide are high achievement and equity [e.g., No Child Left Behind Act in 2001 in the USA; Strategic Framework for Education and Training of the European Union (European Commission 2016)].

The aim for high achievement stems from the evidence that academic achievement is a strong prerequisite for many other effectiveness criteria, e.g., well-being (Samuel et al. 2013), drop-outs in higher education (Bowers et al. 2013), and labour market participation (Heckman et al. 2006). Many studies have led to the conclusion that higher INQUA relates to higher achievement, although most of them were considering achievement in secondary education. Whereas in general, results show positive relations between CM, SC and CM, respectively, on one side and achievement on the other side (e.g., Kunter and Baumert 2006; Klieme et al. 2009), relations between each of the three factors and achievement could not be established in all studies. For instance, Scherer et al. (2016) found strong relations between CM and achievement, with positive and significant relations in all educational systems under study (i.e., Australia, Canada, and America), whereas only weak relations between CA and achievement were found, which were only significant for the Australian sample.

The aim for equity, as stated by Kyriakides and Creemers (2011), 'suggests that differences in outcomes should not be attributable to differences in areas such as wealth, income, power, or possessions' (p. 240). In the current realm of increasing social exclusion, poverty and migration, socio-economic status (SES) and ethnicity are two characteristics of students that are frequently studied (e.g., Lee and Burkam 2002; Farkas 2017). In line with Kyriakides and Creemers' definition (2011), this study assumes that more equity is reached when keeping the relation between achievement and individual background to a minimum. Opposed to the relation between INQUA and achievement, the role of INQUA in the equity debate has far less been studied. The main research question in this debate is whether students coming from different backgrounds are in need of differentiated educational practices. The literature concludes that differential effects are mostly absent, indicating that low versus high SES students, and non-native versus native speakers are in need of the same educational practices to obtain their highest possible

achievement (MacBeath and Mortimore 2001). However, the absence of effective educational practices is most harmful for achievement of more disadvantaged students: ineffective educational practices result in a larger achievement gap between students with less versus more advantaged backgrounds compared to more effective educational practices (Campbell et al. 2004). This indicates that teachers play a vital role regarding equity by assuring high INQUA.

Research aim

Based on findings in the literature, this study aims to contribute to the knowledge base on measuring INQUA and on its role in the international aim for high achievement and equity. The study will accomplish this by investigating the extra items on INQUA in TIMSS 2015 in three educational systems. The three main objectives are: (a) to evaluate the psychometric quality of the extra items in TIMSS 2015 measuring INQUA, (b) to contribute to the validation of the relation between INQUA and achievement across different educational systems, and (c) to investigate the role of INQUA in aiming for equity.

Research design

Research questions

We set forth the following research questions:

RQ1: Do the items reliably measure the three dimensions of INQUA as classroom constructs? And if not, can we build reliable scales with the items, capturing the dimensions of INQUA as classroom constructs?
RQ2: To what extent can INQUA contribute to achievement?
RQ3: To what extent can INQUA contribute to equity? I.e., to what extent does INQUA moderate the relation between student background characteristics and achievement?

Data and sample

We make use of data from TIMSS 2015 Grade 4 gathered in Flanders, Germany and Norway. In Flanders, data was collected from 5404 students in 295 classes from 153 schools. In Germany, 3948 students from 214 classes in 204 schools participated. In Norway, 4164 students from 219 classes in 139 schools participated. Alongside the achievement tests that measured students' math and science ability, student, teacher, principal and parent questionnaires were administered to gather information on students' broader (learning) environments. Instruments were administered in the official language of respective educational systems, i.e., Dutch, German, and Norwegian.

Measures

INQUA

CM was measured by five items that asked students to indicate the frequency of different classroom events on a 4-point Likert-scale ranging from (0) 'never or hardly ever' to (3) 'every lesson'. The scales of SC and CA items consisted of statements which students had to evaluate on a 4-point Likert scale ranging from (0) 'I disagree a lot' to (3) 'I agree a lot'. SC was measured in Flanders and Germany through the use of five items. The SC

scale in Norway (ten items) stated teacher practices which students had to evaluate on a 4-point Likert scale ranging from (0) 'never or hardly ever' to (3) 'every lesson'. The CA scale consisted of seven items.

In this study, CM, SC and CA were taken into account as latent constructs at both the student (STUD CM, STUD SC and STUD CA) and class level (CM, SC and CA).

Student outcome: math achievement (MATH)

MATH is represented by the five plausible values representing students' underlying achievement, made available by TIMSS 2015 (Mean = 500; SD = 100 in 1995). In line with the prescribed use by von Davier et al. (2009), all five plausible values were used in the analyses, using techniques for multiple imputation.

Student background variables

Socio-economic status (SES)   We used books at home as an indicator of socio-economic status (SES),[2] distinguishing five categories, ranging from (1) 0–10 books at home to (5) more than 200 books at home.

Ethnicity   Language spoken at home (LANG), based on students' answers, was taken into account as an indicator of ethnicity,[3] distinguishing four categories ranging from (0) never to (3) always speak language of test at home.

SES and LANG were taken into account as manifest, quantitative variables[4] at the student level. They were taken into account as control variables in RQ2. In RQ3, the relation between student background variables and MATH is investigated as an indication of (in) equity.

Table 1 shows correlations between all variables. As can be noticed, at the between level a significant correlation exists between CM and MATH in all educational systems, whereas no significant correlation between SC and MATH is found. The correlation between CA and MATH is only significant (and negative) in Flanders.

---

[2] Various indicators of SES were investigated in this study before selecting 'books at home' as the SES indicator: The 'Home Resources for Learning' scale made available by TIMSS 2015 (Martin et al. 2016), as well as the different variables constituting this SES indicator (i.e., number of books in the home, number of home study supports, number of children's books in the home, highest level of education of either parent, and highest level of occupation of either parent). All analyses in this study including the construct of SES (RQ2 and RQ3), were repeated with all different SES indicators in the three educational systems. Although regression weights slightly differed in strength, significance of the results regarding (1) the relation between SES and MATH, (2) the relation between the constructs of INQUA and MATH, and (3) the moderating role of INQUA in the relation between SES and MATH, remained the same, regardless of the SES indicator used. The choice for books at home stems from the lowest proportion of missing values in this variable: 1% in Flanders, 17% in Germany, and 2% in Norway (by means of comparison: the 'Home Resources for Learning'-scale showed missing rates of 10% in Flanders, 43% in Germany, and 56% in Norway). Furthermore, analyses revealed that the amount of books at home is a factor which is highly correlated to the 'Home Resources for Learning'-scale (ranging from .72 in Flanders to .75 in Norway).

[3] Language spoken at home, both registered from students and parents, is the only indicator of ethnicity available in TIMSS. Using information from the student questionnaire stems from the lower rate of missingness in this variable: 1% in Flanders, 15% in Germany, and 2% in Norway (compared to a missing rate of 8% in Flanders, 41% in Germany and 56% in Norway based on information from the parent questionnaire).

[4] Two different operationalizations of SES and LANG were investigated: on the one hand taking them into account as dichotomous variables (for SES a distinction was made between students having (1) 0–100 books at home, and (2) more than 100 books at home; for LANG a distinction was made between students (1) never or sometimes speak language of test at home, and (2) almost always or always speak language of test at home), on the other hand taking them into account as continuous variables. The inclusion of both variables as continuous variables in this study stems from two reasons: (1) analyses including SES/LANG in this study (RQ2 and RQ3) were repeated making use of both operationalizations in all educational systems and did not reveal any different conclusion on the significance of the various relations estimated, and (2) the linear relation found between SES/LANG and MATH in preliminary explorative analyses.

Analyses

Multilevel structural equation modelling (SEM) (Nachtigall et al. 2003) was used to answer the research questions. All analyses were performed in MPlus (Muthén and Muthén 2012). We made use of SEM as this enables us to (a) look at the latent variable structure of the dimensions of INQUA (RQ1), excluding the measurement error that is present when investigating the observed students' answers, (b) investigate the relation between the latent variables and the dependent variable (RQ2), and (c) estimate multiple regression equations simultaneously (RQ3) (ibid). Student participants were nested within classes to investigate INQUA as a class construct and to come to terms with the hierarchical structure of the data.[5]

Measurement invariance was partially reached in previous research (Wendt et al. 2016). However, we opted for separate analyses in each educational system in order to investigate the structure of the data more closely in each system. This decision was made because Norway used a different SC scale and because Wendt et al. (2016) reported problems with the data structure of the CA scale across educational systems. This implies that whenever comparisons between educational systems are made in the rest of this article they are descriptive in nature rather than tested in a statistical model.

To answer RQ1, in a first step we calculated intra-class correlation coefficients (ICCs) and performed confirmatory factor analyses (CFA) to investigate to what extent our data provide evidence for the three dimensions of INQUA as classroom constructs. Firstly, ICC[2]s were evaluated as an indication of the extend to with the dimensions of INQUA can be looked at as class-level constructs. Whereas ICC[1] is merely the proportion of variance at the group level (Raudenbush and Bryk 2002), ICC[2] is calculated to evaluate the reliability of the group mean (Bliese 1998), taking the group size *(k)* into account. ICC[2] is defined as (Bliese 1998):

$$\text{ICC}[2] = \frac{k \times ICC[1]}{1 + (k-1) \times ICC[1]}$$

ICC[1] is calculated as:

$$\text{ICC}[1] = \frac{\tau^2}{\tau^2 + \sigma^2}$$

In this formula, τ^2 represents the variance between groups and σ^2 the variance within groups. If not all groups are of the same size, in most cases the mean group size can be chosen (Bliese 2000), which was performed by this study. Bliese (2000) indicated in his simulation study that a group size > 10 reveals reliable estimates. With $\bar{k}_{\text{Flanders}} = 18.0$, $\bar{k}_{\text{Germany}} = 14.4$, and $\bar{k}_{\text{Norway}} = 18.7$, class size averages are high enough to obtain reliable ICC[2]s. An ICC[2] > .60 indicates that the construct can be looked at as a group level construct (Schneider et al. 1998).

Secondly, a CFA model was estimates (see Fig. 1). In this CFA, students with missing data on all scale items were excluded, leading to the exclusion of 82 students in Flanders, 635 students in Germany and 28 students in Norway. Factor loadings and model fit

---

[5] In Flanders, a three-level model with students nested in classes nested in schools showed that variance was situated at only one higher level. As in many cases schools had only one (Grade 4) class, variances at the class and school level cannot sufficiently be disentangled.

Table 1 Correlation of variables

Within

	Flanders					Germany					Norway				
	(1)	(2)	(3)	(4)	MATH	(1)	(2)	(3)	(4)	MATH	(1)	(2)	(3)	(4)	MATH
SES	.11***	.02*	−.01	−.13***	.21***	.13***	.06**	.03*	−.15***	.33***	.13***	.03	.02	−.08**	.24***
LANG (1)		.10***	.07**	−.08***	.14***		.09***	.02	−.04*	.12***		.04	.02	−.06*	.11***
STUD CM (2)			.18***	−.26***	.16***			.16***	−.16***	.16***			.15***	.05	.17***
STUD SC (3)				.22***	.01				.54**	.10***				.73***	.00
STUD CA (4)					−.42***					−.25***					−.16***

Between

	Flanders			Germany			Norway		
	(5)	(6)	MATH	(5)	(6)	MATH	(5)	(6)	MATH
CM	.34***	.08	.30***	.11	.39***	.38***	.53***	.54***	.49***
SC (5)		.90***	.03		.97***	−.03		.83***	.07
CA (6)			−.50***			.06			.08

Correlations were estimated in MPlus by means of a model without any confounders, only including the two variables for which a correlation was estimated. In estimating correlations of between level variables, the variables of interest were also modelled at the within level, as suggested by Marsh et al. (2012)

*** p < .001; ** p < .01; * p < .05

indices (with following cut-off values taken into account to accept the model: CFI > .95, RMSEA < .05, and SRMR < .08; based on the recommendations of Hu and Bentler (1999) were inspected to evaluate whether the current items reliably measure the three dimensions of INQUA as classroom constructs or whether they are in need of adaptation.

Whenever ICC[2]s and results of CFA did not confirm the three-dimensional structure of INQUA at the classroom level (i.e., ICC[2] < .60, factor loadings at the between level with an absolute value < .40, factor loadings with opposite direction to the majority of items, non-significance of factor loadings, and non-acceptable model fit), adaptations of the scales were made based on results of additional exploratory factor analyses (EFA) to capture the factor structure of INQUA in our data. This approach was suggested by Scherer et al. (2016). Based on an iterative process of EFA and CFA, a final CFA model was estimated which best captures the constructs of INQUA at the class level.

A multilevel composite reliability measure (multilevel ω) of the final constructs was calculated, which 'reflects the degree to which group-level differences in a researcher's observed data can be generalized to represent between-group differences in a construct of interest' (Geldhof et al. 2014, p. 75). In line with the recommendation of Geldhof et al. (2014), reliability was estimated with the formula:

$$\omega = \frac{\left(\sum_{i=1}^{k} \lambda_i\right)^2}{\left(\sum_{i=1}^{k} \lambda_i\right)^2 + \left(\sum_{i=1}^{k} \theta_{ii}\right)^2}$$

Here λ represents the factor loading of item i, and θ_{ii} is the unique variance of item i. In line with α, ω 'represents the ratio of a scale's estimated true score variance relative to its total variance' (Geldhof et al. 2014, p. 73). Additionally, it 'acknowledges the possibility of heterogeneous item-construct relations and estimates true score variance as a function of item factor loadings $(\lambda_i)'$ (Geldhof et al. 2014, p. 73). The level-specific parameter estimates were used to calculate the respective within and between composite reliability. Values of .20, .50, and .80 reflect respectively low, medium and high reliability.

The final factor structure found in RQ1 was further extended with structural relations between the latent variables in order to answer RQ2 and RQ3. In RQ2, the relation between the constructs of INQUA as found in RQ1 and MATH was modelled, controlling for SES and LANG. Full information maximum likelihood (FIML) estimation was used to handle missing data.

Figure 2 represents the model tested in answering RQ3. Moderation models were estimated to investigate the extent to which INQUA might counter inequity, facilitated through the means of the relation between SES/LANG and MATH. Due to reasons of complexity, the analyses were done separately for each dimension of INQUA. In looking at the relation between SES and achievement, we controlled for LANG (Fig. 2). Analogously, in investigating the relation between LANG and achievement, we controlled for SES. As FIML is not available in MPlus in random slope moderation models, cases
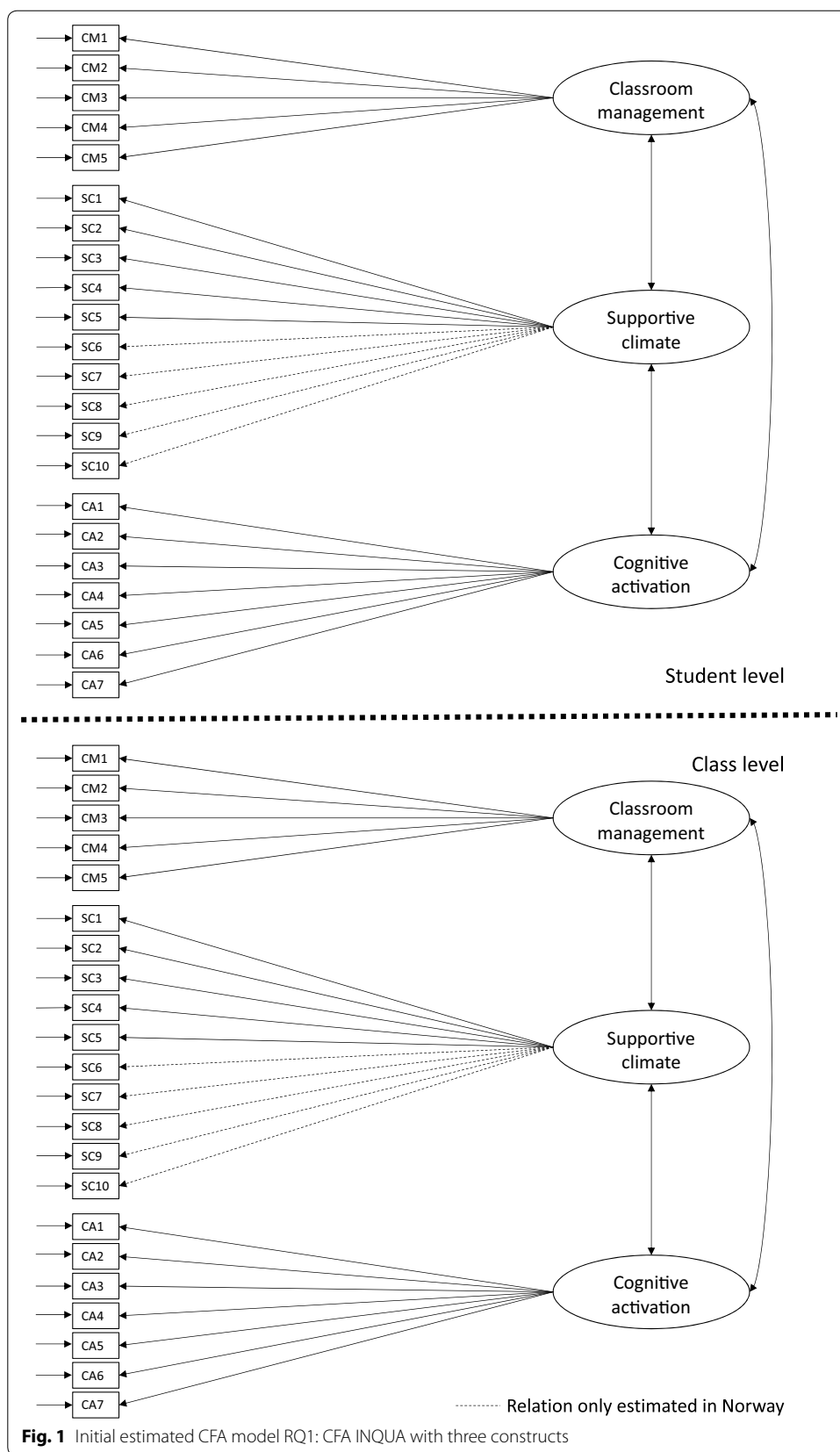
Fig. 1 Initial estimated CFA model RQ1: CFA INQUA with three constructs

with missing scores on the interaction variables were deleted.[6] Factor loadings and significance of the moderation terms included were investigated to answer RQ3. As a starting point, we analyzed the direct relation between student characteristics and MATH (Fig. 3).

In modelling a between-level construct, we always also included, as suggested by Marsh et al. (2012), the latent variable structure of the construct at the within-level to control for bias/noise in the individual students' answers. To account for the unequal probability of classes and students to be selected in the sample due to the (two-stage) sample design of TIMSS, appropriate weights were included in the analyses at both levels.[7]

Results

Results RQ1

ICC[2]s and results of the initial CFA (see Fig. 1) can be found in Table 2. For CM and SC, in all educational systems, ICC[2] is > .60. Results also show high (> .40), significant, and positive factor loadings for all items of both scales at the between level. These results indicate that the used items reliably capture the underlying constructs of CM and SC as class-level constructs in all three educational systems.

In modelling the CA scale some issues were encountered. First of all, some items had to be removed for the CFA model to converge. In Flanders and Germany, this was the case for item 6, an item relating to motivational aspects of math rather than CA by the teacher. In Norway, items 1 and 4 had to be deleted for reasons of model convergence. Both items refer to the degree of task difficulty and in this may be closely linked to CA. Second, ICC[2] is < .60 in all educational systems, indicating problems in the reliability of this construct in representing the group-level mean. Third, items' factor loadings on the between level are low (< .40), non-significant and/or negative for some items (in Flanders and Germany, this holds for items 1, 3 and 4; in Norway, this holds for item 3). Items 2, 5, and 7, however, do show high, positive and significant factor loadings at the between level in the three educational systems. At last, model fit indices show non-acceptable model fit based on the fit indices of CFI and SRMR between.

Based on these results, multilevel EFA were performed to further investigate the data structure. Solutions for three and four factors both at the within and between level were examined using Geomin rotation of the factor loading matrix. First, EFA provided evidence for CM and SC being two distinct factors in all three educational systems, both at the within and between level. Second, EFA confirmed item 6 to be an outlier: factor loadings were non-significant and below .40 on all factors extracted. Third, results of EFA (i.e., model fit and factor loadings) extracting three versus four factors, indicated that the CA scale consists of two subcomponents, i.e., a subcomponent CA(1), consisting of items 1, 3 and 4, and a subcomponent CA(2), consisting of items 2, 5 and 7. There

---

[6] In Table 8 (results of RQ3), the amount of students taken into account in the analyses is indicated. Whereas missingness in Flanders and Norway is very low in SES and LANG, missing data in Germany is higher (see p. 10), implying that caution is needed in interpreting the results.

[7] Based on the recommendations of Rutkowski et al. (2010) weights were calculated and taken into account in the analyses as follows:
Student-level weight = basic student weight * student weight adjustment (WGTFAC3 * WGTADJ3)
Class-level weight = basic class weight * class weight adjustment * basic school weight * school weight adjustment (WGTFAC2 * WGTADJ2 * WGTFAC1 * WGTADJ1).

Fig. 2 Estimated model RQ3
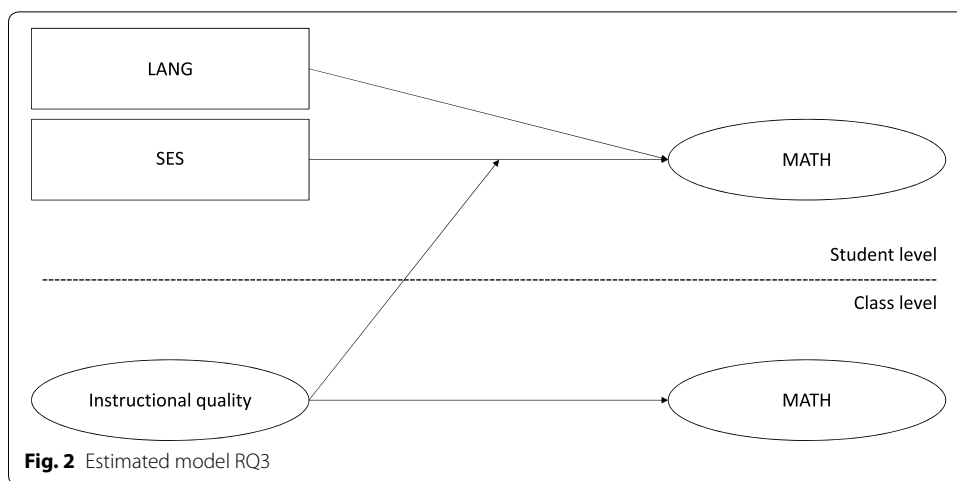
is also a content-driven reason to distinguish between these two components: CA(1) seems to relate closest to issues of CA of students (by means of triggering higher order thinking), whereas CA(2) touches more upon issues of support for learning (to see all CA-items, refer to Appendix). Table 3 shows the correlation of the two components of CA with the other variables included.

However, whereas evidence for this four-factor structure of CM, SC, CA(1) and CA(2) is clearly found at the within level in all educational systems, results at the between level are less explicitly pronouncing the adequacy of this four-factor structure, with factor loadings on the between level of the third and fourth factor showing no clear factors to be distinguished. Based on these results of the EFA, together with the results of the initial CFA model, we estimated three additional CFA models: (a) model A, including all four factors both at the within and between level, (b) model B, including the four-factors structure at the within level, but including at the between level only those constructs which showed an ICC[2] > .60 in Model A, and (c) model C, only including (at both the within and between level) the constructs which showed an ICC[2] > .60 in Model A. Model fit indices of these three estimated CFA models are reported in Table 4. Model A revealed ICC[2]s for CA(1) < .60 in all countries, and CA(2) < .60 in Germany and Norway (not reported in Table 4). Consequently, CA(1) is not included at the between level in Model B and completely left out of the analyses in Model C in all educational systems; and in Germany and Norway, CA(2) is excluded from the between level in Model B and from the within and between level in Model C.

Based on these findings, we conclude that Model C is the model best describing the data structure in all three educational systems. The model is graphically presented in Fig. 4. Results of this final CFA model can be found in Table 5, showing high, positive, and significant factor loadings at the between level, together with acceptable model fit in all educational systems. Table 5 further shows that $\omega_{between}$ > .80 for all constructs included (with an exception of .78 for the CM scale in Norway), indicating high reliability of the constructs modelled.

Fig. 3 Direct link between background characteristics and achievement (basis for moderation models)
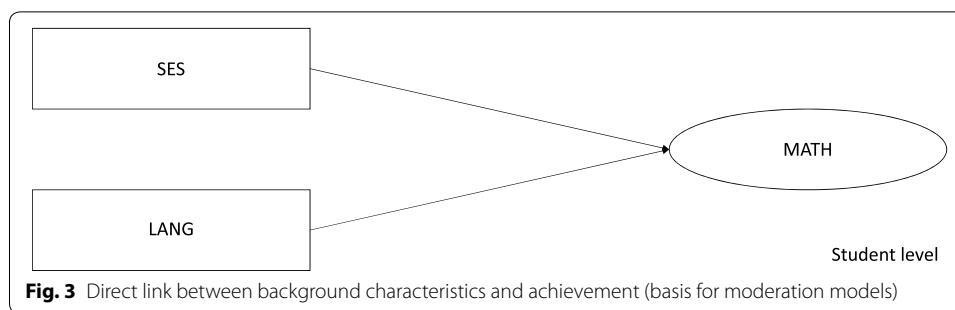
Results RQ2

Based on the result findings of RQ1, Fig. 5 graphically depicts the model at the class level chosen to answer RQ2.

After controlling for student background in SES and LANG, several significant relations between INQUA and MATH are found (Table 6). In Germany and Norway, the relation between CM and MATH is positive ($b_{Germany} = .39$; $p < .01$; $b_{Norway} = .69$, $p < .001$). In Flanders, a significant, positive relation exists between SC and MATH ($b = .72$, $p < .001$) and a significantly negative relation exists between CA(2) and MATH ($b = -1.11$, $p < .001$) indicating that more CA (or more support for learning) in classrooms relates to lower achievement. After student background variables are accounted for, the INQUA scales explain 45, 11 and 73% of the between-level variance in Flanders, Germany, and Norway, respectively. The total proportion of variance explained, taking into account all variables, sums up to 22% in Flanders, 17% in Germany, and 17% in Norway.

Noteworthy are the differences in the regression weights reported in Table 6 and the correlations reported in Table 1, mainly regarding Flanders. Taking into account the confounders of SES and LANG and investigating the INQUA constructs separately, reveals that the differences found are due to the mutual consideration of all INQUA constructs in one model, rather than due to the inclusion of the student level confounders. In Flanders, investigating separately the relation between CM, SC, and CA(2), respectively, and MATH, only controlling for SES and LANG (results only reported in text, not in tables) shows regression weights of $b_{CM} = .28$ ($p < .05$), $b_{SC} = .01$ ($p > .05$), and $b_{CA(2)} = -.64$ ($p < .001$). In Germany and Norway, analyses reveal regression weights of $b_{CM} = .37$ ($p < .01$), and $b_{SC} = -.01$ ($p > .05$), and $b_{CM} = .53$ ($p < .001$), and $b_{SC} = .08$ ($p > .05$), respectively.

Results RQ3

Table 7 shows the results of estimating the direct relation between SES/LANG and MATH as depicted in Fig. 3. In the three educational systems, a significant correlation between students' SES/LANG and MATH is shown. As can be seen in Table 8, the results of the estimated moderation model (Fig. 2) reveal no significant cross-level moderations terms, indicating that the relation between SES/LANG and achievement is not moderated by the different constructs of INQUA investigated.

Table 2 Results initial CFA analyses RQ1

	Flanders	Germany	Norway
	b (SE)	*b (SE)*	*b (SE)*
Model within			
Classroom management			
CM–CM1	.46*** (.02)	.56*** (.02)	.55*** (.02)
CM–CM2	.63*** (.02)	.71*** (.01)	.75*** (.02)
CM–CM3	.61*** (.02)	.71*** (.02)	.70*** (.02)
CM–CM4	.40*** (.02)	.55*** (.02)	.39*** (.03)
CM–CM5	.46*** (.02)	.60*** (.02)	.51*** (.03)
Supportive climate[a]			
SC–SC1	.71*** (.02)	.61*** (.02)	.57*** (.02)
SC–SC2	.68*** (.02)	.65*** (.02)	.54*** (.03)
SC–SC3	.54*** (.02)	.64*** (.02)	.66*** (.02)
SC–SC4	.70*** (.01)	.73*** (.02)	.64*** (.02)
SC–SC5	.39*** (.02)	.41*** (.03)	.63*** (.02)
SC–SC6			.66*** (.02)
SC–SC7			.58*** (.02)
SC–SC8			.64*** (.02)
SC–SC9			.62*** (.02)
SC–SC10			.54*** (.02)
Cognitive activation[b]			
CA–CA1[b]	.74*** (.01)	.76*** (.02)	
CA–CA2	.29*** (.02)	.40*** (.03)	.64*** (.02)
CA–CA3	.76*** (.02)	.78*** (.02)	.44*** (.02)
CA–CA4[b]	.56*** (.02)	.57*** (.02)	
CA–CA5	.28*** (.02)	.36*** (.03)	.61*** (.02)
CA–CA6[b]			.60*** (.02)
CA–CA7	.26*** (.02)	.37*** (.03)	.55*** (.02)
Correlations			
CM–SC	.15***	.15***	.10***
CM–CA	− .26***	− .18***	.06
SC–CA	.18***	.35***	.76***
Model between			
Classroom management			
CM–CM1	.96*** (.03)	.96*** (.02)	.97*** (.03)
CM–CM2	.89*** (.03)	.99*** (.02)	.96*** (.02)
CM–CM3	.89*** (.03)	.94*** (.02)	.93*** (.03)
CM–CM4	.81*** (.06)	.92*** (.04)	.97*** (.05)
CM–CM5	.83*** (.05)	.92*** (.03)	.91*** (.04)
Supportive climate[a]			
SC–SC1	.83*** (.05)	.88*** (.06)	.97*** (.06)
SC–SC2	.94*** (.04)	.88*** (.07)	.94*** (.08)
SC–SC3	.99*** (.04)	.94*** (.05)	.88*** (.07)
SC–SC4	.81*** (.06)	.93*** (.04)	.89*** (.06)
SC–SC5	.75*** (.07)	.89*** (.08)	.97*** (.07)
SC–SC6			.97*** (.06)
SC–SC7			.97*** (.16)
SC–SC8			.91*** (.07)
SC–SC9			.81*** (.18)
SC–SC10			.50*** (.18)

Table 2 (continued)

	Flanders	Germany	Norway
	b (SE)	*b (SE)*	*b (SE)*
Cognitive activation[b]			
CA–CA1[b]	.29 (.25)	.23** (.08)	
CA–CA2	.93*** (.05)	1.01*** (.07)	.99*** (.11)
CA–CA3	.38 (.24)	− .36 (.85)	.38 (.26)
CA–CA4[b]	.20 (.23)	.19 (.24)	
CA–CA5	.98*** (.08)	.81*** (.09)	.90*** (.09)
CA–CA6[b]			.92*** (.07)
CA–CA7	.74*** (.08)	.83*** (.10)	.82*** (.10)
Correlations			
CM–SC	.33***	.18	.56***
CM–CA	.08	.35**	.64***
SC–CA	.83***	.96***	.82***

Model fit			
RMSEA	.04	.05	.03
CFI	.83	.85	.93
SRMR within	.08	.08	.04
SRMR between	.21	.23	.12
Chi^2 (df)	2168.35*** (202)	1948.74*** (202)	1443.72*** (334)

ICC	ICC[1]	ICC[2]	ICC[1]	ICC[2]	ICC[1]	ICC[2]
CM	.40	.92	.24	.82	.24	.86
SC	.13	.74	.10	.62	.14	.75
CA[b]	.00	.00	.00	.00	.05	.48

*** p < .001; ** p < .01

[a] As stated in the paragraph 'Measures' under the heading of 'Research design', in Norway supportive climate was measured by different items. The items and accompanying numbering of items in the analyses of Norway stem from the item numbering in Appendix

[b] For the model to converge, in Flanders and Germany, item 6 had to be deleted; in Norway, item 1 and 4 had to be deleted

Discussion

Taking results of this study into account, this section will endeavour to further expand upon discussions of (a) measuring INQUA, and (b) the contribution of INQUA to educational outcomes. We conclude this discussion with (c) our overall reflections on the concept of INQUA and its dimensions.

Measuring INQUA: possibilities and challenges

Our findings can serve as input in the future development of valid and reliable measures of INQUA and its dimensions. This study made use of a data-driven approach to obtain the best possible model of INQUA and its relation to educational outcomes, which implies that caution is needed in generalizing the results. However, the same model was found to be valid across three different data-sets and educational contexts which strengthens our conclusions to a great extent.

Whereas the construct of CM was validly and reliably established in this study, we can relate to the literature findings in that some issues are noted in the constructs of SC and

Table 3 Correlation of the two components of CA with the other variables

	Flanders		Germany		Norway	
	CA(1)	CA(2)	CA(1)	CA(2)	CA(1)	CA(2)
Within						
SES	− .16***	− .08**	− .17***	− .02	− .12***	− .03
LANG	− .10***	− .09**	− .04	− .01	− .05***	− .06*
STUD CM	− .27***	− .06*	− .18***	.01	− .09**	.12***
STUD SC	.10***	.65***	.23***	.76***	.38***	.79***
STUD CA(1)		.41***		.51***		.62***
MATH	− .42***	− .19***	− .28***	− .05	− .24***	− .07*
Between						
CM	− .56***	− .01	− .18***	− .02	− .06	.54***
SC	− .04	.70***	.00	.84***	.37	.79***
CA(1)		.55***		.55**		.54**
MATH	− .81***	− .64***	− .48***	− .13	− .35*	− .04

*** p < .001; ** p < .01; * p < .05

Table 4 Model fit indices for different CFA models estimated

	Model A	Model B	Model C
Flanders			
RMSEA	.03	.03	.02
CFI	.93	.93	.96
SRMR within	.04	.04	.03
SRMR between	.10	.21	.07
Germany			
RMSEA	.03	.03	.03
CFI	.95	.94	.97
SRMR within	.03	.03	.02
SRMR between	.10	.30	.06
Norway			
RMSEA	.03	.03	.03
CFI	.93	.93	.95
SRMR within	.03	.03	.03
SRMR between	.12	.33	.08

CA (see Praetorius et al. 2014). First of all, regarding the SC scale, a content-wise distinction of the scale in Flanders and Germany versus Norway should be noted, with items measuring student–teacher relationships versus learning support activities. Nevertheless, both scales were seen to be valid and reliable in the respective countries. Second, the items measuring the underlying CA scale were found to be problematic. We showed that further differentiation in the CA scale is needed to establish a valid and reliable construct, with one component referring to issues of support for learning (CA(2)), another capturing the triggering of higher order thinking processes of students (CA(1)). Splitting the CA scale made it possible to pinpoint to some extent the overlap between CA and SC, pointing to the thin line between the support dimensions of both concepts: we found rather low correlations between CA(1) with SC, and high correlations between CA(2) and SC.

Fig. 4 Final CFA model RQ1: CFA INQUA with CA split in two components
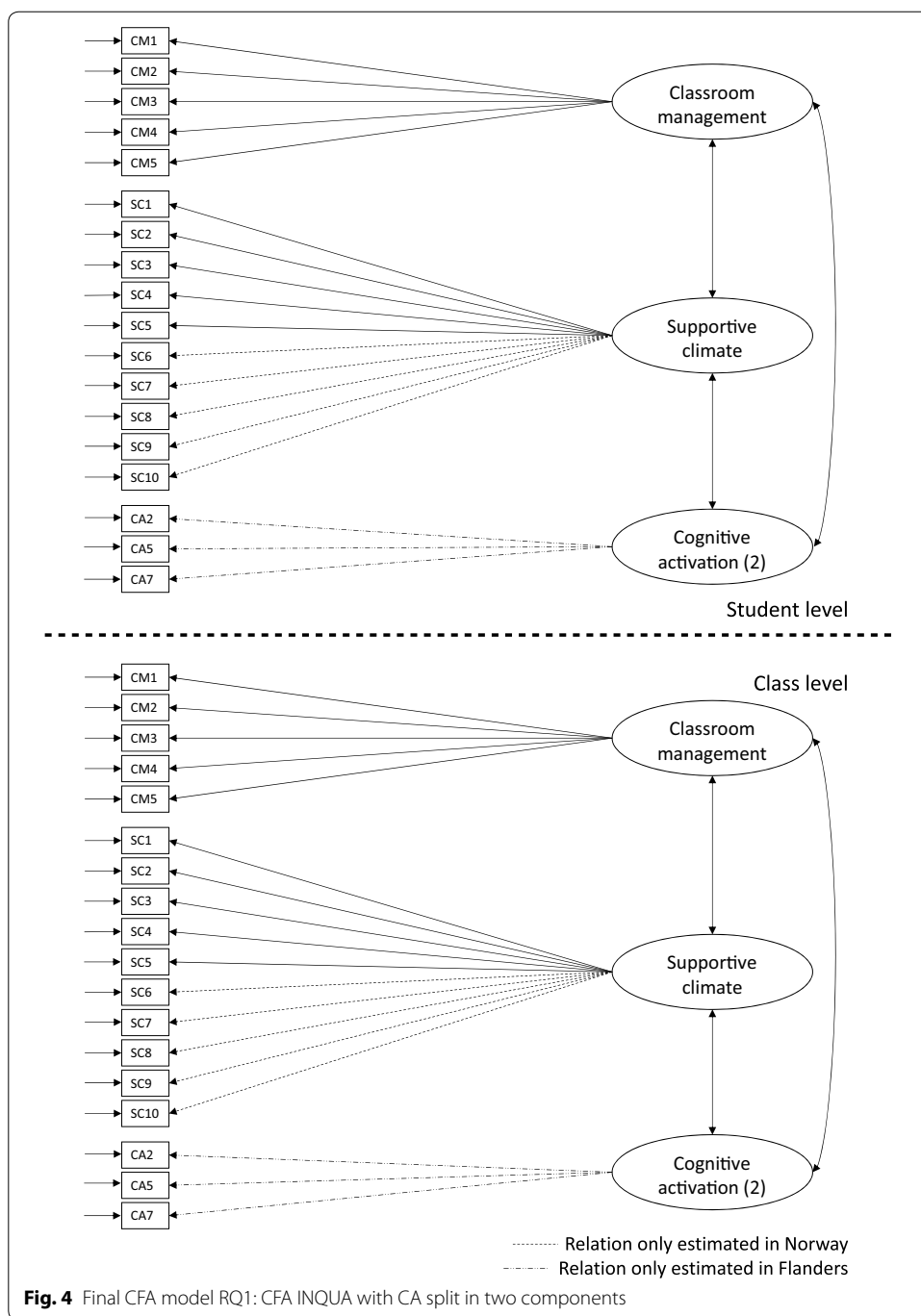
Second, our results question the premise that the construct of 'INQUA' is a teacher characteristic. In contrast to Fauth et al. (2014) our results suggest that student ratings are not merely reliable measures of INQUA as a class construct. Furthermore, whereas previous research indicated that teachers' CM and SC are stable across contexts, the literature suggests that CA is highly context-specific and is influenced heavily by the group and the subject taught (Fauth et al. 2014). That said, and taking into account the issues with the CA scale of this study, we seek to question to what extent INQUA, and more specifically CA, are also student-specific constructs, and furthermore, to what extent

Table 5 Results final CFA analyses RQ1 with adapted scale constructs

	Flanders[b]	Germany	Norway
	b (SE)	b (SE)	b (SE)
Model within			
Classroom management			
CM–CM1	.54*** (.02)	.58*** (.02)	.59*** (.02)
CM–	.51*** (.03)	.71*** (.02)	.70*** (.02)
CM–CM3	.46*** (.03)	.68*** (.03)	.63*** (.02)
CM–CM4	.43*** (.03)	.52*** (.02)	.39*** (.03)
CM–CM5	.47*** (.02)	.57*** (.02)	.52*** (.03)
Supportive climate[a]			
SC–SC1	.69*** (.02)	.60*** (.03)	.57*** (.02)
SC–SC2	.68*** (.02)	.65*** (.02)	.53*** (.03)
SC–SC3	.57*** (.03)	.63*** (.02)	.66*** (.02)
SC–SC4	.69*** (.02)	.74*** (.02)	.64*** (.02)
SC–SC5	.40*** (.02)	.41*** (.03)	.60*** (.02)
SC–SC6			.65*** (.02)
SC–SC7			.58*** (.02)
SC–SC8			.66*** (.02)
SC–SC9			.59*** (.02)
SC–SC10			.53*** (.02)
Cognitive activation component 2			
CA–CA2	.60*** (.03)		
CA–CA5	.52*** (.02)		
CA–CA7	.45*** (.03)		
Correlations			
CM–SC	.18***	.16***	.11***
CM–CA(2)	− .04		
SC–CA(2)	.63***		
Model between			
Classroom management			
CM–CM1	1.00*** (.03)	.96*** (.03)	1.00*** (.03)
CM–CM2	.83*** (.04)	.99*** (.03)	.91*** (.04)
CM–CM3	.85*** (.05)	.94*** (.03)	.87*** (.06)
CM–CM4	.79*** (.05)	.91*** (.04)	.96*** (.05)
CM–CM5	.83*** (.04)	.91*** (.03)	.92*** (.05)
Supportive climate[a]			
SC–SC1	.85*** (.05)	.88*** (.06)	.97*** (.06)
SC–SC2	.94*** (.04)	.92*** (.06)	.94*** (.09)
SC–SC3	.95*** (.05)	.85*** (.07)	.88*** (.07)
SC–SC4	.83*** (.05)	.96*** (.04)	.87*** (.06)
SC–SC5	.75*** (.07)	.90*** (.08)	.98*** (.07)
SC–SC6			.98*** (.06)
SC–SC7			.96*** (.15)
SC–SC8			.90*** (.08)
SC–SC9			.80*** (.19)
SC–SC10			.52*** (.14)
Cognitive activation component 2			
CA–CA2	.87*** (.07)		
CA–CA5	.87*** (.10)		
CA–CA7	.65*** (.11)		

Table 5 (continued)

	Flanders[b]		Germany		Norway	
	b (SE)		*b (SE)*		*b (SE)*	
Correlations						
CM–SC	.36***		.12		.54***	
CM–CA(2)	.04					
SC–CA(2)	.69***					
Model fit						
RMSEA	.02		.03		.03	
CFI	.96		.97		.95	
SRMR within	.03		.02		.03	
SRMR between	.07		.06		.08	
Chi^2 (df)	501.99*** (120)		249.47*** (64)		706.42*** (170)	
ICC	**ICC[1]**	**ICC[2]**	**ICC[1]**	**ICC[2]**	**ICC[1]**	**ICC[2]**
CM	.36	.91	.23	.81	.22	.84
SC	.16	.77	.10	.62	.14	.75
CA(2)	.18	.79				
ω	**Within**	**Between**	**Within**	**Between**	**Within**	**Between**
CM	.64	.99	.78	.99	.73	.78
SC	.76	.99	.76	.99	.85	.92
CA(2)	.55	.99				

*** p < .001

[a] As stated in the paragraph 'Measures' under the heading of 'Research design', in Norway supportive climate was measured by different items. The items and accompanying numbering of items in the analyses of Norway stem from the item numbering in Appendix

[b] For reasons of uniformity of the models across countries, we also estimated a model in Flanders not including the construct of CA(2). This model shows similar fit indices and does not significantly differ from the model reported with regard to factor loadings of CM and SC

researchers will be able to establish a CA construct that goes beyond individual perceptions. Or, should we look at these 'perceptions' as realities, in other words, do teachers give differential INQUA to different students in one class group—whether or not consciously? To extend our knowledge with regard to this issue, further research is required concerning measurement of INQUA, including investigating data from multiple sources.

INQUA in relation to outcomes of educational systems: between dream and reality?

A main contribution of this study is that we looked at INQUA in relation to math achievement across three countries. Our results give some insights in (a) differences across educational systems in the contribution of INQUA to achievement and equity, and (b) challenges in the contemporary educational practices across educational systems.

Firstly, some relations between INQUA and achievement were established for different educational systems in line with stated hypotheses in the background section above, providing further evidence for which factors relate to student achievement. First, the positive relation found between CM and math achievement in all educational systems in this study further sustains the hypothesized and previously empirically established relation (e.g., Kunter and Baumert 2006; Klieme et al. 2009; Scherer et al. 2016) between

Fig. 5 Estimated model RQ2
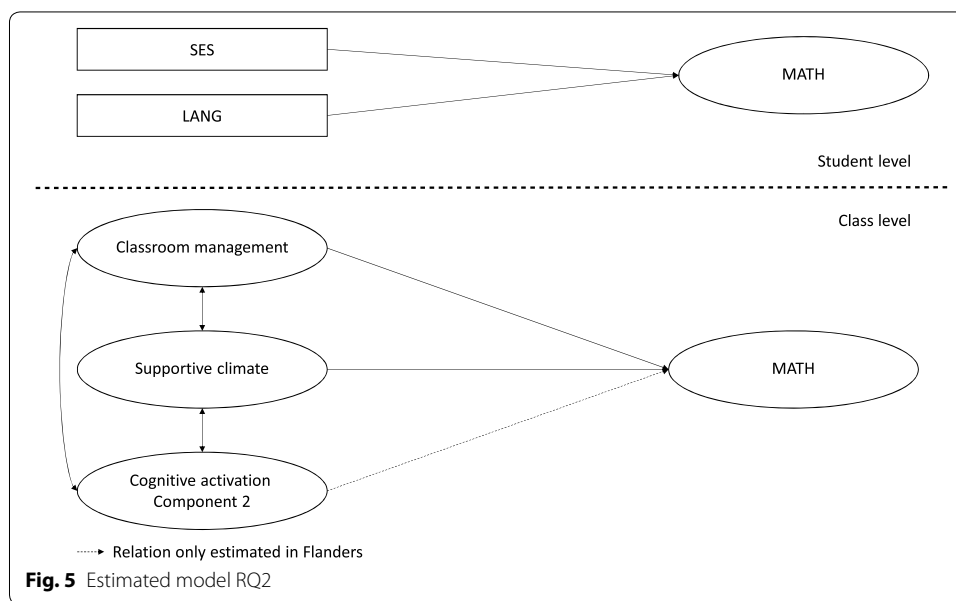
both concepts. Our results, again, stress the importance of good classroom management in obtaining high students' achievement. Second, whereas the positive relation found in earlier research between SC and math (e.g., Kunter and Baumert 2006; Klieme et al. 2009) is established in this study for Flanders (after controlling for a.o. CM), no evidence for this relation is found in the other two countries. Third, in line with results of Scherer et al. (2016), no evidence is found for a positive relation between CA and achievement, opposed to the findings of Kunter and Baumert (2006) and Klieme et al. (2009). In contrast, the negative relation between support for learning (CA(2)) and math in Flanders contradicts our hypothesis of a positive correlation between those variables. As we did not take into account prior achievement, reversed causality might explain this negative relation. Considering this assumption, our results might point to the fact that high achieving students in Flanders do not receive sufficient support for learning, resulting in reduced chances for the highest possible achievement.

However, this study is limited in the sense that we only took into account student background factors as confounders in investigating the relation between INQUA and achievement. First of all, although sensitivity analyses with regard to SES and LANG were performed which showed stable results across operationalization of both variables, one needs to be aware of the possible bias introduced in the results by treating ordinal variables as though they were continuous in this study. Furthermore, INQUA, is hypothesized to be influenced by a complex set of input and process factors, which may also affect the relation between INQUA and achievement. It is for instance assumed that a diverse set of teacher quality characteristics (pedagogical content knowledge, professional beliefs, work-related motivation, self-regulation aspects, etc.) lead to INQUA (Berliner 2005; Blömeke et al. 2016). Taking these variables into account might give a more differentiated picture of INQUA in relation to achievement. In this, further research is needed, investigating the intertwined relation of INQUA with different educational practices and context factors, and investigating how this complex interplay between different factors might benefit students' achievement. Additionally, the

Table 6 Results analyses RQ2

	Flanders	Germany	Norway
	b	*b*	*b*
Model within			
MATH on SES	.23***	.34***	.24***
MATH on LANG	.14***	.09***	.09***
Model between			
MATH on CM	.04	.39**	.69***
MATH on SC	.72**	− .08	− .28
MATH on CA(2)	− 1.11***		
Correlations			
CM–SC	.36***	.13	.55***
CM–CA(2)	.03		
SC–CA(2)	.67***		
Model fit[a]			
RMSEA	.03	.04	.03
CFI	.90	.91	.93
SRMR within	.05	.05	.04
SRMR between	.09	.07	.09
Chi^2 (df)	1068.86*** (170)	789.57*** (103)	1006.39*** (229)

*** $p < .001$; ** $p < .01$

[a] Whereas fit indices CFI and SRMR between show an insufficient model fit, all model fit indices indicate acceptable model fit when only constructs showing a significant relation with MATH are taken into account (i.e., SC and CA(2) in Flanders, and CM in Germany and Norway). On a related note, models only taking into account CM, or only CM and CA(2) in Flanders also both show acceptable model fit

Table 7 Results direct relation between background characteristics and achievement

	Flanders	Germany	Norway
	b	b	b
Model within			
MATH on SES	.20***	.32***	.23***
MATH on LANG	.12***	.08***	.08***
Model fit			
Loglikelihood (df)	− 21,350.84 (4)	− 14,435.61 (4)	− 17,130.67 (4)

*** $p < .001$

alternating results in the relation of CM and SC with achievement in Flanders, depending on taking into account the constructs of INQUA separately versus together, also points to the intertwined relation between the constructs of INQUA mutually. Apparently, in Flanders, the relation between CM and MATH is explained by the SC offered in classrooms, whereas this is not the case in Germany and Norway. Further investigating the intertwined relations between the constructs of INQUA is needed to further sustain our conceptual model of INQUA.

These issues of reversed causality, confounding factors, and intertwined relations between the constructs of INQUA also point to the main limitation of this study. The cross-sectional character of the data left us with merely correlational conclusions rather than establishing causal relations between INQUA and achievement. Although, in this study, we assumed causal hypotheses and research questions, based on our research

Table 8 Results analyses RQ3

	Flanders						Germany				Norway			
	b	b	b	b	b	b	b	b	b	b	b	b	b	b
Model within														
MATH on SES	.19***	.19***	.19***	.19***	.19***	.19***	.32***	.32***	.32***	.32***	.22***	.22***	.22***	.22***
MATH on LANG	.12***	.12***	.12***	.12***	.12***	.12***	.08***	.08***	.07***	.08***	.08***	.08***	.08***	.08***
Model between														
MATH on CM	.39***			.37***			.47**		.51**		.57***		.58***	
MATH on SC	.07	.02			.02		.03		.06		−.01		.04	
MATH on CA(2)			− 1.25***			− 1.20***								
Cross-level moderation														
MATH on SES * CM	−.01						.07				−.19			
MATH on SES * SC		−.07						−.13				.00		
MATH on SES * CA(2)			−.10											
MATH on LANG * CM				−.02					−.08				−.05	
MATH on LANG * SC					.02					.03				−.21
MATH on LANG * CA(2)						.12								
N	5321	5321	5323	5323	5323	5323	3284	3284	3330	3330	4091	4091	4091	4095
Model fit														
Loglikelihood (df)	− 55,364.77 (31)	− 55,175.35 (31)	− 42,667.00 (23)	− 55,405.92 (31)	− 55,198.73 (31)	− 42,659.97 (23)	− 33,393.38 (31)	− 33,764.04 (31)	− 33,725.83 (31)	− 34,081.40 (31)	− 42,748.42 (31)	− 43,441.46 (31)	− 42,763.18 (31)	− 43,385.81 (31)

*** p <.001; ** p <.01

design we cannot claim causal inferences between INQUA and achievement from the results found. In light of this correlational design, our operationalization of equity (the non-existence of a relation between background variables and achievement) might be questioned. Not controlling for initial ability implies that the correlation, rather than the causal influence between background variables and achievement, is investigated. As such, the correlation between student background variables and achievement might be explained by differences in, amongst others, educational and/or societal preferences and aspirations amongst children. This, in turn, may lead to variation in achievement, an effect which has been observed frequently. On a related note, whereas one may question the operationalization of books at home and language spoken at home in our analyses, sensitivity analyses took into account different variables as well as different operationalization of the variables and did show rather robust results which strengthen the results of our analyses.

Secondly, the results of the moderation analyses provide valuable input for the equity debate. Despite the large sample sizes and consequently high statistical power, no moderation effects were found to be significant. The absence of moderation effects is in line with literature findings indicating that all students benefit equally from the same effective educational practices (MacBeath and Mortimore 2001), indicating that teachers do not have to set up differential educational practices for students coming from different social and ethnic background.

The concept of INQUA and its dimensions: one size fits all?

Based on our findings regarding the measuring of INQUA and its relation to educational outcomes, we may conclude that the research field on INQUA is still in need of further cross-cultural conceptual clarification of, and differentiation between the dimensions at stake.

Whereas this study provides further evidence for the definition and construct of CM, this study shows that efforts should focus on disentangling the issues at stake to capture the SC and CA dimensions. In this, we may raise the question, 'to what extent can the entire learning process be captured within (merely) three dimensions. Moreover, is it possible and desirable to include several key aspects of learning, (i.e., giving support for learning to students, activating students' thinking processes, etc.), in one dimension, namely CA. Or, should we extend the number of dimensions in order to facilitate conceptual clarification and differentiation between dimensions (in particular between SC and CA), which might facilitate investigating and capturing the effects of educational practices? An alternative might be given by the Bill & Melinda Gates Foundation (2010), in which seven dimensions of effective educational practices are differentiated. Next to the concepts of control (in line with CM) and care (in line with SC), effective educational practices are captured in five dimensions by the Foundation that capture the different steps in the student learning process: clarifying, challenging, captivating, conferring and consolidating.

As only partial measurement invariance could be established in previous research on the same data (Wendt et al. 2016), this study is limited in the sense that educational systems were not investigated in a joint model, restricting the conclusions to a descriptive comparison across educational systems rather than to statistical testing of differences. However, this may point out a more important question, i.e., 'to what extent can

we establish a unified, cross-cultural concept of INQUA and its underlying dimensions?' Whereas Scherer et al. (2016) found measurement invariance of INQUA and its three factor structure across three countries, they point to the high similarity between the educational systems under study. The absence of measurement invariance in our data, together with several findings of our study (i.e., differences between countries in (a) evidence of the constructs to be classroom constructs, (b) the relation between INQUA and math achievement, and (c) the large differences in the variance explained by INQUA), point to the need for further research on the cross-cultural validation of INQUA and its dimensions. Next to the need to establish a general framework of INQUA, research should also investigate the differential structure of INQUA and its plausible differential effect on educational outcomes across different educational systems.

Conclusion

This study showed that INQUA might serve as a catalyst for practitioners aiming towards high achievement for their students. However, issues in the conceptualization and cross-cultural measurement of INQUA are raised, questioning the extent to which the three well-known dimensions of INQUA (a) are well-defined and might be sufficiently differentiated from each other, (b) sufficiently capture the diverse set of educational practices relating to students' educational outcomes and (c) can be established across countries in a unified manner.

Author details
[1] Centre for Educational Effectiveness and Evaluation, Faculty of Psychology and Educational Sciences, KU Leuven, Dekenstraat 2, Box 3773, 3000 Leuven, Belgium. [2] Faculty of Psychology and Educational Sciences, KU Leuven, Campus Kulak Kortrijk, Etienne Sabbelaan 51, Box 7800, 8500 Kortrijk, Belgium. [3] Faculty of Educational Sciences, Psychology and Sociology, Institute for School Development Research, TU Dortmund University, Emil-Figge-Straße 50, 44227 Dortmund, Germany. [4] Department of Teacher Education and School Research, University of Oslo, Postboks 1099, Blindern, 0317 Oslo, Norway.

Appendix

See Table 9.

Table 9 Extra items on instructional quality in TIMSS 2015

Scale	Item
Classroom management	Question: How often do these things happen in your mathematics lessons? Response categories: never or hardly ever (0), some lessons (1), most lessons (2), every lesson (3)
	(CM1) Students don't listen to what the teacher says
	(CM2) There is noise and disorder
	(CM3) Our mathematics teacher has to wait a long time for students to quiet down
	(CM4) Students cannot work well
	(CM5) Students don't start working for a long time after the lesson begins
Supportive climate Flanders and Germany	Question: To what extent do you agree with the following statements? Response categories: disagree a lot (0), disagree a little (1), agree a little (2), agree a lot (3)
	(SC1) Our mathematics teacher is nice to me even when I make a mistake
	(SC2) Our mathematics teacher cares about me
	(SC3) Our mathematics teacher tells me how to do better when I make a mistake
	(SC4) Our mathematics teacher likes me
	(SC5) Our mathematics teacher believes that I can solve difficult tasks
Supportive climate Norway	Question: How often does your mathematics teacher do the following? Response categories: never or hardly ever (0), some lessons (1), most lessons (2), every lesson (3)
	(SC1) Show an interest in every student's learning
	(SC2) Give extra help when students need it
	(SC3) Help students with their learning
	(SC4) Continue teaching until the students understand
	(SC5) Give students the opportunity to express opinions
	(SC6) Set clear learning goals
	(SC7) Explain what he/she expects us to learn
	(SC8) Ask questions to check if we understand the content of the lesson
	(SC9) Explain how new topics relate to previous lessons, at the beginning of the lesson
	(SC10) Summarize what we have covered at the end of the lesson
Cognitive activation	Question: To what extent do you agree with the following statements? Response categories: disagree a lot (0), disagree a little (1), agree a little (2), agree a lot (3)
	(CA1) In our mathematics lessons, we are working on tasks that I have to think about very thoroughly
	(CA2) Our mathematics teacher asks me what I have understood and what I haven't
	(CA3) Our mathematics teacher asks questions that I have to think about very thoroughly
	(CA4) Our mathematics teacher gives us tasks that seem to be difficult at a first glance
	(CA5) Our mathematics teacher asks what we know about a new topic
	(CA6) Our mathematics teacher gives us tasks I like to think about
	(CA7) Our mathematics teacher wants me to be able to explain my answers

Publisher's Note

References

Berliner, D. (2005). The near impossibility of testing for teacher quality. *Journal of Teacher Education, 56,* 205–213.

Bill & Melinda Gates Foundation (2010). *Learning about teaching. Initial findings from the measures of effective teaching project*. Retrieved from https://graphics8.nytimes.com/packages/pdf/education/20101211education.pdf.

Bliese, P. D. (1998). Group size, ICC values, and group-level correlations: A simulation. *Organizational Research Methods, 1*(4), 355–373.

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 349–381). San Francisco, CA: Jossey-Bass.

Blömeke, S., Olsen, R. V., & Suhl, U. (2016). Relation of student achievement of the quality of their teachers and instructional quality. In T. Nilsen & J. E. Gustafsson (Eds.), *Teacher quality, instructional quality and student outcomes. Relationships across countries, cohorts and time* (pp. 21–50). Cham: Springer International Publishing.

Bowers, A. J., Sprott, R., & Taff, S. A. (2013). Do we know who will drop out?: A review of the predictors of dropping out of high school—Precision, sensitivity, and specificity. *High School Journal, 96*(2), 77–100.

Campbell, J., Kyriakides, L., Muijs, D., & Robinson, W. (2004). *Assessing teacher effectiveness: Developing a differentiated model*. London: Routledge.

European Commission. (2016). *Strategic framework for education and training.* Retrieved from http://ec.europa.eu/education/policy/strategic-framework/index_en.htm.

Farkas, G. (2017). *Human capital or cultural capital? Ethnicity and poverty groups in an urban school district*. London: Routledge.

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.

Ferguson, R. F. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan, 94*(3), 24–28.

Geldhof, G. J., Preacher, J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods, 19*(1), 72–91.

Heckman, J. J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics, 24*(3), 411–482.

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

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

Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research, 9,* 231–251.

Kyriakides, L., & Creemers, B. P. (2011). Can schools achieve both quality and equity? Investigating the two dimensions of educational effectiveness. *Journal of Education for Students Placed at Risk, 16,* 237–254.

Lee, V. E., & Burkam, D. T. (2002). *Inequality at the starting gate: Social background differences in achievement as children begin school*. Washington, D.C.: Economic Policy Institution.

MacBeath, J., & Mortimore, P. (2001). *Improving school effectiveness*. Buckingham: Open University Press.

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., et al. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist, 74,* 106–124.

Martin, M. O., Mullis, I. V. S., Foy, P., & Hooper, M. (2016). *TIMSS 2015 international results in Mathematics.* Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/timss2015/international-results/.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 international results in mathematics.* Retrieved from Boston College, TIMSS & PIRLS International study center website, via http://timssandpirls.bc.edu/timss2015/international-results/.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus version 7.0* [computer software]. Retrieved from https://www.statmodel.com/index.shtml.

Nachtigall, C., Kroehne, U., Funke, F., & Steyer, R. (2003). (Why) should we use SEM? Pros and cons of structural equation modeling. *Psychological Research online, 8*(2), 1–22.

Nilsen, T., & Gustafsson, J. E. (2016). *Teacher quality, instructional quality and student outcomes. Relationships across countries, cohorts and time.* Cham: Springer International Publishing.

O'Dwyer, L. M., Wang, Y., & Shields, K. A. (2015). Teaching for conceptual understanding: A cross-national comparison of the relationship between teachers' instructional practices and student achievement in mathematics. *Large-scale Assessments in Education, 3*(1), 1–30.

OECD. (2005). *PISA 2003: Technical report*. Paris: OECD Publishing.

OECD. (2013). *PISA 2012: Assessment and analytical framework. Mathematics, reading, science, problem solving and financial literacy*. Paris: OECD Publishing.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38,* 109–119.

Praetorius, A.-K., Lenske, G., & Helmke, A. (2012). Observer ratings of INQUA: Do they fulfill what they promise? *Learning and Instruction, 22*(6), 387–400.

Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of INQUA across lessons. *Learning and Instruction, 31,* 2–12.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: Sage.

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Recommendations for secondary analysis and reporting. *Educational Researcher, 39*(2), 142–151.

Samuel, R., Bergman, M. M., & Hupka-Brunner, S. (2013). The interplay between educational achievement, occupational success, and well-being. *Social Indicators Research, 111*(1), 75–96.

Scherer, R., Nilsen, T., & Jansen, M. (2016). Evaluating individual students' perceptions of instructional quality: An investigation of their factor structure, measurement invariance, and relations to educational outcomes. *Frontiers in Psychology, 7,* 110. https://doi.org/10.3389/fpsyg.2016.00110.

Schneider, B., White, S. S., & Paul, M. C. (1998). Linking service climate and corner perceptions of service quality: Test of a causal model. *Journal of Applied Psychology, 83,* 150–163.

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). *What are plausible values and why are they useful? IERI Monograph Series Issues and Methodologies in Large-Scale Assessments.* Hamburg: IER Institute, Educational Testing Service.

Wendt, H., Nilsen, T., Kasper, D., & Van Damme, J. (2016). *Student ratings of INQUA: How valid are they across countries?* In European Conference on Educational Research (ECER). Dublin, Ireland, 22–26 August 2016.

Wubbels, T., Brekelmans, M., & Hooymayers, H. P. (1992). Do teacher ideals distort the self-reports of their interpersonal behavior? *Teaching and Teacher Education, 8,* 47–58.