Reviewing Measurement Invariance of Questionnaire Constructs in Cross-National Research: Examples from ICCS 2016

Wolfram Schulz

Australian Council for Educational Research (ACER)

wolfram.schulz@acer.edu.au

Paper prepared for the Annual Meeting of the American Educational Research Association,
Washington D.C., 8 – 12 April 2016

Reviewing Measurement Invariance of Questionnaire Constructs in Cross-National Research: Examples from ICCS 2016

Introduction

International studies face the challenge of obtaining comparable measures across diverse national contexts, which are characterized by differences with regard to many factors such as language, culture, and educational contexts. In addition, in many countries there may also be differences across sub-national contexts (for example across regions or language groups). International studies in the field of education are increasingly incorporating ways of assessing the measurement invariance of the data obtained from tests or questionnaire used in these programs. In particular with regard to questionnaire items, there is a growing body of research indicating that questionnaire formats such as Likert-type rating scale items may not measure respondents' beliefs consistently across diverse cultural or linguistic contexts.

This paper describes how issues of measurement invariance were reviewed based on data from the international field trial of the IEA Civic and Citizenship Education Study (ICCS 2016), which studies the ways young people in lower secondary education are prepared to undertake their roles as citizen in wide range of countries (see Schulz, Ainley, Fraillon, Losito, & Agrusti, 2016; Schulz, Fraillon, Ainley, Losito & Kerr, 2008; Schulz, Ainley, Fraillon, Kerr & Losito, 2010). The survey includes tests and questionnaire to assess students' cognitive and non-cognitive learning outcomes as well as contextual variables.

The field trial analyses incorporated explorative and confirmatory factor analysis, as well as covariances (multiple-group modelling) and item response theory (IRT) which specifically focused on a review of measurement invariance. This paper compares the different methodologies used for analysis, presents examples from the analyses of ICCS 2016 field trial data, and discusses implications for future research on measurement invariance of questionnaire data in cross-national studies.

Theoretical background

In comparative international studies the source measurement instruments tend to be the same for all countries, but each country collects data using adapted and translated versions of the source instruments. Research has shown that differences between source and target language can lead to differences in item responses (see Harkness, Pennell, & Schoua-Glusberg, 2004; Mohler, Smith, & Harkness, 1998).

Van de Vijver and Tanzer (1997) concluded that instruments might work properly but characteristics of cultural groups of respondents may introduce bias in measurement while Byrne (2003) distinguished three different kind of bias due to cultural differences:

- *Construct bias* refers to cases where a construct may be meaningful in one country, but not another country;
- *Method bias* refers to cases where data are biased by differences in responses to the instruments caused by cultural traits; and
- *Item bias* refers to bias that occurs at the level of the individual item.

International studies related to educational research have established sophisticated quality assurance procedures to ensure a maximum of comparability with regard to the adaptation and translation of source instruments in participating countries (in the case of ICCS, see Malak, Yu, Schulz, & Friedman, 2011). However, concerns about the comparability of measures, in particular with regard to those derived from questionnaire persist (see Heine, Lehman, Peng & Greenholtz, 2002; Schulz, 2009; Schulz & Fraillon, 2011; van de gaer, Grisay, Schulz & Gebhardt, 2012).

Typically, questionnaires are used to measure latent variables through student responses to sets of items (e.g. agreement or disagreement with a series of statements) which reflect the construct of interest) and can be scaled to derive summary variables. To be able to compare the resulting measures in cross-national studies it is essential that the underlying measurement model is the same or at least highly similar in each participating country. In principle, measurement invariance is confirmed if individuals with the same score on the same measurement instrument have the same standing on the underlying construct that is measured, independently of the sub-group (here national sample) they belong to. If this assumption does not hold, invalid conclusions may be drawn from comparisons across national contexts.

There are different ways of reviewing the level of measurement invariance. Within the framework of structural equation modelling of variance-covariance structures, multiple-group analyses can be conducted with comparisons of model fit for the same factor structure across models with different levels of parameter constraints (see Little, 1997; Little & Slegers, 2005; Meredith, 1993; Sörbom, 1974). Here, the focus typically lies on a review of invariance for the overall measurement model. Within the context of Item Response Theory (see Rasch, 1960; Hambleton, Swaminathan, & Rogers, 1991), the review of measurement invariance is referred to as differential item functioning (DIF), which consists of finding different item parameters within different sub-samples (see Hambleton, & Rodgers, 1995; Perrone, 2006). Here, attention is mainly focused on the invariance of item parameters across sub-samples (in our case: countries).

The paper will illustrate how construct validity and measurement invariance for student questionnaire scales were assessed during the field trial analyses for ICCS 2016. An application of thorough analyses of measurement equality at this stage of a cross-national study provides information to the final selection process for the main survey, where preference may be given to item sets which not only have satisfactory psychometric quality for the pooled international sample, but also tend to have similar measurement characteristics across participating countries. It should be noted, however, that differences in measurement may also be seen as informative and that the decision about retaining item material should not only be driven by psychometric criteria.

With regard to its research questions, firstly, the paper will attempt to review to what extent different ways of assessing measurement invariance (i.e. multiple-group modelling based on variance-covariance structures and item response modelling) provide similar information about the extent to which scales derived from questionnaire items are comparable across different national contexts. Secondly, it will review to what extent item sets have different levels of measurement invariance depending on whether their content is more or less influenced by national context variables. For example, student attitudes toward different national institutions are expected to vary more across countries than student views regarding their confidence to engage in different forms of participation.

Data and methods

Data

The analyses presented in this paper were based on results from the ICCS 2016 field trial which was conducted between October and December 2014. National field trial instruments were adapted and translated at national centres and underwent adaptation reviews, translation verification (by independent language experts) and final layout verification by the International Study Centre. Data were collected, using a 45-minute test of civic knowledge and a 40-minute international student questionnaire, from 19,090 students in their eighth year of schooling from 470 lower-secondary schools in 20 different countries.[1] Furthermore, school principals and random samples of 15 teachers at each sampled schools provided data about school-level contexts for civic and citizenship education.

The purpose of the field trial was to trial the procedures developed for survey administration as well as reviewing the appropriateness and psychometric quality of the instruments developed for assessing students, teachers and school principals. The analysis presented in this paper focus on data from the international student questionnaire. In order to trial a larger pool of items, three overlapping questionnaire forms were administered which ensured that all possible combinations of item sets and questions could be analysed.

The analyses will focus on three different item sets included in the ICCS 2016 field trial student questionnaire:

- *Trust in civic institutions*: Students were asked to rate their trust in different institutions
- *Citizenship self-efficacy*: Students rated their confidence in undertaken different form of civic engagement
- *Civic participation at school*: Students were asked to report on the extent to which they had participated in different forms of student participation at school.

It should be noted that field trial samples were not very large (typically about 1000 students from 25 schools) and that for each item set responses were collected from only about two thirds of these students due to the questionnaire design. Therefore, it would not be appropriate to infer to the underlying populations. Therefore, country-level results are only presented without identifying participating countries.

The field trial analyses of item sets designed to derive scales included the following steps:

- *Analyses of missing responses*: This aspect was reviewed for all questions and items in the questionnaires.
- *Exploratory factor analyses*: At the first stage, exploratory factor analyses (without assumptions about factor structure) were conducted.
- *Confirmatory factor analyses*: At the second stage, confirmatory factor analyses (making assumptions about factor structure informed by the exploratory analyses) were undertaken, followed by multiple-group analyses to review the invariance of the measurement models.

[1] In four countries the field trial was administered at a later stage and their data could not be included in the international data analysis.

- *Review of item statistics and reliabilities*: At national and international levels, (adjusted) item-total correlations as well as Cronbach alpha reliabilities were computed and reviewed.
- *Correlations with test scores of civic knowledge*: Given that item responses in many of the questions are expected to be influenced by student knowledge about civic issues, correlations with test scores were routinely reported.
- *Item response modelling*: For each scaled item set, Rasch Partial Credit models were estimated once for the pooled international field trial sample and in a second step with interaction parameters to review measurement invariance at the item level.

The analyses presented in this paper will focus on the exploratory and confirmatory factor analyses, a review of scale reliabilities and test score correlations, as well as IRT analyses. While exploratory factor analyses for the pooled sample are presented to provide insight in the overall measurement characteristics, scale reliabilities and scale correlations with civic knowledge at the country level will illustrate consistency of the derived scale across national contexts. The main focus, however, will be on the review of measurement invariance through multiple-group models (with different constraints) and country-item interactions as provided by an IRT analysis.

Classical Item and Scale Analysis

Cronbach's alpha coefficient (Cronbach, 1951) provides an estimate of the internal consistency of each scale, and is probably one of the most commonly used estimates of scale reliability. While there are no agreed criteria regarding the level of acceptable internal consistency, values over 0.7 are typically viewed as satisfactory and values over 0.8 as "highly reliable" (see for example, Nunnally & Bernstein, 1994). For the analyses of field trial questionnaire data Cronbach's alpha coefficients were reported but interpreted within the broader context of other indicators, such as the correlations between individual items and the score derived from all other items in a scale (adjusted item-total correlations).

Correlations with civic knowledge (Pearsons' product-moment coefficients) were used to indicate to which extent item responses and the scales derived from them corresponded to the level of knowledge and understanding students had regarding civic and citizenship issues. Even though there are no agreed interpretations, coefficients below 0.1 were viewed as "unsubstantial", above 0.2 as "moderate" and above 0.5 as "strong".

Exploratory Factor Analysis

Exploratory factor analysis (EFA) is based on the analysis of the variance- and covariance structure of items. It and is typically used at preliminary analyses stages to review expected dimensionality of questionnaire items both within and across countries, often in the analyses of field trial data in large-scale assessments.

Given the typically categorical nature of questionnaire items, it is generally recommended to apply factor analysis that uses appropriate estimation methods, such as for example weighted likelihood estimation with tetrachoric or polychoric correlations. Software packages like MPLUS (Muthén, & Muthén, 2012), which was used for the ICCS 2016 field trial analysis, offer procedures that allow conducting EFA for categorical variables.

Within the context of the field trial analyses for ICCS 2016, results from the EFA were used to examine the factor loadings of items forming a scale within countries and across countries. In this way, for example, an examination of the factor loadings for the constituent items on a construct may reveal that a particular item may not contribute to a construct than other and, hence, it might be

suggested as a candidate to be dropped from the questionnaire in the main survey. The structures which emerge from exploratory factor analyses can subsequently be confirmed using confirmatory factor analysis or item response modelling.

Confirmatory Factor Analyses

Confirmatory Factor Analysis (CFA) can be carried out by using structural equation modelling (SEM) techniques (see Kaplan, 2000). Within the SEM framework latent variables are linked to observable variables via measurement equations: An observed variable x is defined as

$$x = \Lambda_x \xi + \delta \ ,$$

where Λ_x is a q x k matrix of factor loadings, ξ denotes the latent variable(s) and δ is a q x 1 vector of unique error variances.

The expected covariance matrix is fitted according to the theoretical factor structure. Model estimates can be obtained through minimising the differences between the expected (*) and the observed covariance matrix (S). Measures for the overall fit of a model then are obtained by comparing the expected Σ matrix with the observed S matrix. If the differences between both matrices are close to zero, then the model "fits the data", if differences are rather large the model "does not fit the data".

Model fit was assessed using the Root-Mean Square Error of Approximation (RMSEA), the Comparative Fit Index (CFI) and the Non-normed Fit Index (NNFI) (see Bollen and Long, 1993). RMSEA values over 0.10 are usually interpreted as a sign of unacceptable model fit whereas values below 0.05 indicate a close model fit. RMR values should be less than 0.05. Both CFI and NNFI are bound between 0 and 1 and values between 0.90 and 0.95 indicate a marginally acceptable model fit, with values greater than 0.95 indicating a close model fit.

In international studies, the parameters may vary across country and it may not be appropriate to assume the same factor structure for each population. One way of looking at invariance of factor structures is to use separate CFA within countries and review model fit within each population across countries. This provides insights into the validity of assuming the same factor structure. The estimation of multiple-group models provides a direct test of parameter invariance.

To test parameter invariance, multiple-group modelling as an extension of CFA offers an approach to test the equivalence of measurement models across sub-samples (Little, 1997; Byrne, 2008). If one considers a model where respondents belong to different groups indexed as g = 1, 2, ..., G, the multiple-group factor model becomes

$$x_g = \Lambda_{xg} \xi_g + \delta_g$$

A test of factorial invariance (H_Λ) where factor loadings are defined as being equal (often referred to as “metric equivalence” (Horn & McArdle, 1992) can be defined as

$$H_\Lambda : \Lambda_1 = \Lambda_1 = \Lambda_2 = ... = \Lambda_g$$

Model-fit indices can be compared across different multiple-group models, each with an increasing degree of constraints, from relaxed models with no constraints through to constrained models with largely invariant model parameters. Constraints may be placed on factor loadings, intercepts, factor variances as well as covariances.

In this paper, three different multiple-group models are presented each with different levels of constraints on the parameters in a confirmatory factor analysis:

A. Unconstrained models with all parameters treated as country-specific (*configural invariance*);

B. Models with constrained factor loadings across countries (*metric invariance*);

C. A model with constraints on factor loadings and intercepts (*scalar invariance*)

The last model is the only one which ensures absolute comparability of measurement models, and thus scale scores, across participating countries. When comparing model fit across the three conditions, it needs to be acknowledged that with data from large samples, as is typically the case in international large-scale assessments, even very small differences appear to be significant, and that therefore hypothesis testing using tests of significance tends to be problematic. Therefore, in the model comparisons we have focused on a review of relative model fit (RMSEA, NNFI and CFI) across the three models instead of relying on tests of statistical significance. All confirmatory factor analyses were conducted using the software package MPLUS (Muthén, & Muthén, 2012).

Item Response Theory

For the field trial analysis IRT (Item Response Theory) model is used as a scaling methodology for both the Civics and Citizenship competencies and questionnaire data (see Hambleton, Swaminathan, & Rogers, 1991). The One-parameter (Rasch) model (Rasch 1960) was applied that predicts the probability of selecting the a response to an item on a latent trait θ_n. All IRT analyses were undertaken using ACER ConQuest (Wu, Adams, Wilson, & Haldane, 2007).

For dichotomous items with categories scored 0 and 1, a response with a value of 1 is modelled as

$$P_i(\theta) = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)}$$

where $P_i(\theta_n)$ is the probability of person *n* to score 1 on item *i*. θ_n is the estimated latent trait of person n and δ_i the estimated location of item *i* on this dimension. For each item, item responses are modelled as a function of the latent trait θ_n.

In the case of items with more than two categories, which is typically the case in questionnaire items, the model can be generalised to the so-called Partial Credit Model (Masters, & Wright, 1997) as

$$P_{x_i}(\theta) = \frac{\exp \sum_{k=0}^{x} (\theta_n - \delta_i + \tau_{ij})}{\sum_{h=0}^{m_i} \exp \sum_{k=0}^{h} (\theta_n - \delta_i + \tau_{ij})} \qquad x_i = 0,1,\ldots,m_i$$

where $P_{xi}(\theta_n)$ is the probability of person n to score x on item i. θ_n denotes the person's latent trait, the item parameter δ_i gives the location of the item on the latent continuum and τ_{ij} is an additional step parameter.

The goodness of fit for individual items can be determined by calculating a (weighted) Mean Square Statistic (Wright, & Masters, 1982). Values greater than 1 show that the item is less discriminating than expected by the model, whereas values below 1 indicate a discrimination that is higher than expected. However, it needs to be noted that this type of residual-based statistics needs to be interpreted with caution and only in conjunction with other item fit indicators (see Rost, & von Davier, 1994).

Tests of parameter invariance across national sub-samples can be conducted by calibrating questionnaire items separately within countries and then comparing model parameters and item fit

across countries. However, it is also possible to estimate group effects directly by including further parameters as facets in the IRT scaling model. For the partial credit model, which is typically used as IRT model for scaling questionnaire data, so-called item-by-country interactions can be estimated with the following facet model:

$$P_{x_i}(\theta) = \frac{\exp \sum_{j=0}^{x_i} (\theta_n - (\delta_i - \eta_c + \lambda_{ic} + \tau_{ij}))}{\sum_{h=0}^{m_i} \exp \sum_{j=0}^{h} (\theta_n - (\delta_i - \eta_c + \lambda_{ic} + \tau_{ij}))} \qquad x_i = 0,1,2,\ldots,m_i$$

To obtain estimates of parameter equivalence across a group of national sub-samples c, an additional parameter for national effects on the item parameter λ_{ic} (the item-by-country interaction) is added to the model. The model requires the inclusion of the overall national effect (η_c) in the model to obtain proper estimates. Both item-by-country interaction estimates (λ_{ic}) and overall country effects (η_c) are each constrained to having a sum of 0. Item-by-country interactions (λ_{ic}) can be summarised at the item level by computing the median of absolute values as well presenting the minimum and maximum of the (original) parameters.

As with multiple-group models, given the typically large sample size tests of statistical significance for item parameter variation are not appropriate to assess the acceptability of measurement model variance. For the purpose of reviewing item-by-country interaction (or country DIF) it was assumed that parameters above 0.3 logits indicated DIF.

Results

Trust in Institutions

The ICCS 2016 field trial student questionnaire included a question with 10 international and three optional ICCS 2009 items asking students to rate their trust in civic institutions and groups ("completely", "quite a lot", "a little", "not at all"). Given that function and corresponding views of civic institutions and groups may be heavily influenced by specific national contexts (historical, social, political), measurement models for (a) derived scale(s) are expected to vary.

Table 1 shows the results from an exploratory factor analysis: the two-factor solution had a marginally poor model fit (RMSEA = 0.113, RMR = 0.051) but it confirmed that the first six items had relatively strong loadings on the first factor while the two items related to media and social media loaded on a second factor. Item I, J and K loaded on the first factor but had rather weak factor loadings while item L did not load on either factor. For further scaling analyses only the first six items were used to derive a scale reflecting trust in civic institutions in ICCS 2009 (S_ITRUST).

Table 1 Trust in groups and institutions: Explorative factor analysis

Item	Item wording	Factors 1	2
A	The <national government> of <country of test>	0.86	-0.13
B	The <local government> of your town or city	0.76	-0.02
C	Courts of justice	0.74	-0.03
D	The police	0.63	0.05
E	Political parties	0.81	-0.04
F	<National Parliament>	0.87	-0.11
G	Media (television, newspapers, radio)	-0.02	0.73
H	Social media (e.g. Twitter, blogs etc.)	-0.24	0.87
I	<The Armed Forces>	0.44	0.19
J	Schools	0.40	0.26
K	The United Nations	0.53	0.15
L	People in general	0.23	0.33
EV		**5.18**	**1.48**

* EFA results based on WLSMV estimation with PROMAX rotation (factor loadings for two-factor solution).

A one-dimensional CFA showed poor model fit which was mainly due to considerable residual correlations between items C (courts of justice) and D (police), as well as between E (political parties) and F (national parliament). Table 2 shows the model fit for a CFA which included correlated error terms for these pairs of items, which were estimated at 0.31 between items C and D, and 0.36 between E and F. The model fit was satisfactory for the pooled sample.

Table 2 Trust in civic institutions: Confirmatory factor analysis and multiple-group model comparison

	Overall	Configural	Metric	Scalar
RMSEA	0.079	0.082	0.099	0.139
CFI	0.99	0.99	0.98	0.94
TLI	0.99	0.99	0.98	0.96

* Models estimated using WLSMV estimator.

When comparing the three multiple-group models, the fit was only marginally acceptable for the model with metric invariance and no longer satisfactory for the most constrained (scalar) model. These results suggest that student responses regarding their trust in different institutions, as expected, follow different patterns across different national contexts.

Table 3 Trust in civic institutions: Scale reliabilities and correlations with civic knowledge

Country	Cronbach's alpha	Correlation with civic knowledge
CNT1	**.82**	0.00
CNT2	**.85**	**-0.21**
CNT3	**.86**	0.00
CNT4	**.79**	-0.14
CNT5	**.85**	**0.25**
CNT6	**.83**	**-0.21**
CNT7	**.84**	0.07
CNT8	**.86**	0.18
CNT9	**.83**	0.01
CNT10	**.83**	-0.03
CNT11	**.85**	-0.12
CNT12	**.88**	-0.19
CNT13	**.80**	0.01
CNT14	**.84**	0.18
CNT15	**.86**	0.14
CNT16	**.82**	**-0.23**
CNT17	**.84**	0.01
CNT18	**.87**	0.10
CNT19	**.83**	0.14
CNT20	**.86**	-0.02
Average	**.84**	0.00

* Reliabilities > 0.7 and correlations < -0.20 or > 0.20 in **bold**.

Table 3 records the scale reliabilities for the scale reflecting students' trust in the civic institutions of their country, as well as its correlations with civic knowledge test scores. The reliabilities were satisfactory in all countries but there were no consistent correlations between this scale and civic knowledge test scores. While in some countries there were moderate negative correlations, in others there were weak to moderate positive correlations.

Table 4 Trust in civic institutions: IRT results

Item	Parameter	Weighted mean square (fit) statistic	Item-by-country interaction		
			Median of absolute values	Minimum	Maximum
A: National government	-.12	.93	.23	-1.24	.95
B: Local government	-.12	.98	.15	-.54	.54
C: Courts of justice	-.34	1.02	.28	-.50	.46
D: Police	-.51	1.23	.32	-1.13	1.23
E: Political parties	.84	.96	.16	-.29	.50
F: Parliament	.25	.95	.20	-.49	.86

* ACER Conquest estimates.

Table 4 shows the results from an IRT analysis: All items except D (police) had satisfactory item fit statistics. The item parameters indicate that while courts and police on average attracted higher

levels of trust (i.e. it was easiest to express trust), political parties were the least trust institution (i.e. it was most difficult to trust them). However, the analysis of item-by-country interaction showed that considerable variation of item parameters across countries, in particular for item D ("Police") but also for items C ("Courts of Justice") and A ("National Government"). This corresponds to findings from the multiple-group analysis, which showed considerable lack of measurement invariance across countries.

Students' sense of Citizenship Self-Efficacy

The ICCS 2016 field trial student questionnaire included a set of six items asking students to rate their confidence in taking part in different activities reflecting citizenship engagement ("very well", "fairly well", "not very well", "not at all well"). The item set is expected to have more consistent student responses across countries given that the perceived difficulty of the forms of engagement does not necessarily depend on particular national contexts.

Table 5 Students' citizenship self-efficacy: Exploratory factor analysis

Item		Factors 1	2
A	Discuss a newspaper article about a conflict between countries	0.70	
B	Argue your point of view about a controversial political or social issue	0.73	
C	Stand as a candidate in a <school election>	0.74	
D	Organise a group of students in order to achieve changes at school	0.73	
E	Follow a television debate about a controversial issue	0.68	
F	Write a letter or email to a newspaper giving your view on a current issue	0.69	
G	Speak in front of your class about a social or political issue	0.71	
	Eigen Values	**3.97**	**0.77**

* EFA results based on WLSMV estimation with PROMAX rotation (factor loadings for one-factor solution).

Table 5 shows the results from an exploratory factor analysis: the one-factor solution had only poor model fit (RMSEA = 0.132, RMR = 0.055), however, all items had consistently strong factor loadings. Further CFA revealed that the poor model fit was mainly due to residual correlations (i.e. associations not explained by the latent trait) between items A (discussing a newspaper article) and B (arguing a point of view), and between items C (standing as candidate) and D (organising a group of students).

Table 6 shows the model fit for a unidimensional CFA with estimated residual correlations between the two item pairs (estimated at 0.33 between items A and B, and 0.35 between items C and D). The model fit was excellent for the pooled sample. Across the multiple-group models it was found that while there was not much difference between the unconstrained model (configural) and the model with constrained factor loadings (metric), the most constrained (scalar) model had a somewhat less satisfactory model fit.

Table 6 Citizenship self-efficacy: Confirmatory factor analysis and multiple-group model comparison

	Overall	Configural	Metric	Scalar
RMSEA	0.049	0.072	0.075	0.095
CFI	0.99	0.98	0.98	0.95
TLI	0.99	0.98	0.98	0.97

* Models estimated using WLSMV estimator.

Table 7 records the scale reliabilities for a scale reflecting students' sense of citizenship self-efficacy (S_CITEFF, all seven items), as well as its correlations with civic knowledge. The reliabilities were satisfactory in all field trial countries. While in a number of countries there were weak to moderate positive associations between S_CITEFF and civic knowledge test scores, in many others no (or weak negative) correlations were recorded.

Table 7 Citizenship self-efficacy: Reliabilities and correlations with civic knowledge

COUNTRY	Cronbach's alpha	Correlation with civic knowledge
CNT1	**.80**	-0.03
CNT2	**.84**	0.10
CNT3	**.85**	0.00
CNT4	**.80**	-0.12
CNT5	**.84**	0.10
CNT6	**.74**	-0.04
CNT7	**.85**	0.08
CNT8	**.86**	**0.25**
CNT9	**.80**	**0.21**
CNT10	**.86**	0.14
CNT11	**.83**	0.14
CNT12	**.85**	-0.10
CNT13	**.84**	-0.04
CNT14	**.85**	0.02
CNT15	**.86**	0.17
CNT16	**.77**	-0.05
CNT17	**.80**	-0.06
CNT18	**.84**	**0.23**
CNT19	**.86**	**0.21**
CNT20	**.88**	-0.10
Average	**.83**	0.06

* Reliabilities > 0.7 and correlations < -0.20 or > 0.20 in **bold**.

Table 8 Citizenship self-efficacy: IRT results

Item	Parameter	Weighted mean square (fit) statistic	Item-by-country interaction		
			Median of absolute values	Minimum	Maximum
A: Discuss newspaper article	-.11	1.03	.14	-.46	.35
B: Argue point of view	-.20	.98	.20	-.36	.49
C: Stand as candidate	.10	.99	.12	-.42	.36
D: Organise student group	-.23	1.00	.26	-.47	.53
E: Follow TV debate	.15	1.02	.20	-.55	.40
F: Write to newspaper	.12	1.03	.16	-.45	.43
G: Speak in front of class	.16	1.01	.10	-.30	.30

* ACER Conquest estimates.

Table 8 shows the results of an IRT-based analysis of this item set. All items had satisfactory item fit statistics and there was only limited evidence of item-by-country interactions across national field trial samples. Item D had the highest level of item-by-country interactions which may be due to the greater variation in forms of student organisation across national contexts.

Student Participation at School

Students were asked about their past participation in a number of different school activities ("yes, in the last twelve months", "yes, but more than a year ago", "no"). Given that contexts for school participation are expected to vary across national contexts, it is assumed that only limited measurement invariance would be found for this particular item set.

Table 9 shows the results of the EFA for the school participation items. The one-factor solution provides a good fit (RMSEA = 0.054, RMR = 0.045). However, item H (sports activities) has very low loading and did not scale as well with the other items. Therefore, this item was not included in the scale reflecting student participation at school (S_PRTSCH).

Table 9 Student participation at school: Exploratory factor analysis

		Factors	
Item	**Item wording**	**1**	**2**
A	Voluntary participation in school-based music or drama activities outside of regular classes	0.50	
B	Active participation in an organised debate	0.55	
C	Voting for <class representative> or <school parliament>	0.58	
D	Taking part in decision-making about how the school is run	0.71	
E	Taking part in discussions at a <student assembly>	0.70	
F	Becoming a candidate for <class representative> or <school parliament>	0.66	
G	Participating in an activity to make the school more <environmentally friendly>	0.55	
H	Participating in school-based sports activities outside regular classes	0.39	
	Eigen values	**3.36**	**0.95**

* EFA results based on WLSMV estimation with PROMAX rotation (factor loadings for one-factor solution).

Table 10 Student participation at school: Confirmatory factor analysis and multiple-group model comparison

	Overall	**Configural**	**Metric**	**Scalar**
RMSEA	0.056	0.065	0.136	0.134
CFI	0.97	0.97	0.81	0.76
TLI	0.96	0.95	0.79	0.80

* Models estimated using WLSMV estimator.

Table 10 shows that a one-factorial CFA had reasonable model fit for the pooled sample. However, when comparing multiple-group models with different constraints the model had poor model fit for the more constrained (metric and scalar) models. This suggests that there was considerable measurement variance across participating countries.

Table 11 Student participation at school: Reliabilities and correlations with civic knowledge

Country	Cronbach's alpha	Correlation with civic knowledge
CNT1	**.74**	**0.20**
CNT2	**.78**	0.07
CNT3	**.70**	0.06
CNT4	.65	0.10
CNT5	.68	0.12
CNT6	**.73**	0.08
CNT7	**.73**	0.10
CNT8	.70	0.20
CNT9	.58	0.04
CNT10	**.77**	0.18
CNT11	.70	0.16
CNT12	**.75**	-0.02
CNT13	.68	0.13
CNT14	.65	0.18
CNT15	**.73**	0.16
CNT16	.70	0.05
CNT17	**.70**	0.16
CNT18	**.73**	**0.23**
CNT19	**.74**	0.19
CNT20	.67	0.19
Average	**.70**	0.13

* Reliabilities > 0.7 and correlations < -0.20 or > 0.20 in **bold**.

Table 11 records the reliabilities and the correlations with civic knowledge for the scale reflecting students' past participation at school (S_PRTSCH). The reliabilities were satisfactory in most countries and there were weak to moderate positive correlations with civic knowledge in a majority of countries.

Table 12 Student participation at school: IRT results

Item	Item parameter	Weighted mean square (fit) statistic	Item-by-country interaction		
			Median of absolute values	Minimum	Maximum
A: Music or drama	-.28	1.06	.18	-.68	.98
B: Organised debate	.10	1.04	.30	-.75	1.28
C: Voting	-.99	1.02	.36	-.89	.90
D: Decision-making	.38	.93	.26	-.53	.57
E: Discussions	.35	.96	.14	-1.93	.59
F: Becoming candidate	.31	.97	.24	-.61	.67
G: Environmental action	.13	1.04	.25	-.46	.70

* ACER Conquest estimates.

Table 12 illustrates the results the IRT scaling analysis for the scale S_PRTSCH. All item parameters had satisfactory item fit. In particular for items B ("Active participation in an organised debate") and

C ("Voting for class representative or school parliament") there were a relatively high levels of item-by-country interactions. This corresponds to findings from the multiple-group analysis indicating a lack of measurement invariance across participating countries.

Conclusion

Results from the field trial analyses showed varying degrees of measurement invariance depending on the nature of the questionnaire items as well as with regard to national contexts. They also show that results from a review using multiple-group analysis and an IRT-based review of differential item functioning have very similar results. This is not surprising given that item response models can be conceptualised and are mathematically equivalent to logistic confirmatory factor analyses (see Glöckner-Rist, & Hoijtink, 2003). However, while factor analytic approached (based on the analysis of covariance structures) typically assesses the overall fit of dimensional models, item response modelling is more focused on the performance of individual items. Both approaches can be viewed as complimentary and useful for assessing construct validity in cross-national research, in particular when already applied at earlier stages of international comparative studies.

As expected, scales based on items where responses may be more influenced by national context are less likely to show measurement invariance. Both item sets measuring students' trust in civic institutions and student participation at school presented higher levels of measurement variance across countries, the IRT results showed DIF also for particular items. For the item set measuring students' confidence in their abilities to engage in civic engagement there was a more acceptable level of consistency in measurement models across sub-samples.

The results from this paper suggest that measurement models derived from questionnaire data tend to present a certain lack of measurement invariance, in particular in cases where item responses may be more influenced by contextual factors at the national level. In particular in a study of civic and citizenship education, there is considerable diversity across countries with regard to many variables of interest in this field of research. Therefore, focusing only on constructs and variables that are highly similar in terms of measurement may be problematic and lead to rather narrow scope in comparative studies. The question is also at what point lack of measurement invariance becomes problematic and leads to problematic bias in cross-national surveys. The fact that in recent years there has been a considerable increase in attention paid to the cross-national validity of survey outcomes, gives hope that further research will provide further insight into these issues.

References

Bollen, K. A., & Long J. S. (1993). Testing structural equation models. Newbury Park, CA: Sage.

Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. Psicothema, 20(4), 872-882.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297-334.

Glöckner-Rist, A., & Hoijtink, H. (2003). The best of both worlds: Factor analysis of dichotomous data using item response theory and structural equation modeling. Structural Equation Modeling: A Multidisciplinary Journal, 10(4), 544–565.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of Item Response Theory. Newbury Park, CA: Sage.

Harkness, J., Pennell, B., & Schoua-Glusberg, A. (2004). Survey questionnaire translation and

assessment. In J. Presser, M. Rothgeb, J. Couper, E. Lessler, E. Martin, & E. Singer (Eds.), Questionnaire development evaluation and testing methods (pp. 453–473). Hoboken, NJ: Wiley.

Heine, S. J, Lehman, D. R., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales? The reference group effect. Journal of Personality and Social Psychology, 82(6), 903–918.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. Experimental Aging Research, 18(3-4), 117–144.

Kaplan, D. (2000). Structural equation modeling: foundation and extensions. Thousand Oaks: SAGE publications.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. Multivariate Behavioral Research, 32(1), 53–76.

Little, T. D., & Slegers, D. W. (2005). Factor analysis: Multiple groups. Encyclopedia of statistics in behavioral science. Vol. 2, 617 – 623. Chichester, UK: John Wiley & Sons, Ltd.

Masters, G. N., & Wright, B. D. (1997). The Partial Credit Model. In: van der Linden, W. J. and Hambleton, R. K. (Eds.). Handbook of Modern Item Response Theory (pp. 101-122). New York/Berlin/Heidelberg: Springer.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. Psychometrika, 58(4), 525–543.

Mohler, P. P., Smith, T. W., & Harkness, J. A. (1998). Respondent's ratings of expressions from response scales: A two-country, two-language investigation on equivalence and translation. In J. A. Harkness (Ed.), ZUMA-Nachrichten spezial No.3: Cross-cultural survey equivalence (pp. 159–184). Mannheim: ZUMA.

Muthén, L.K., & Muthén, B.O. (2012). Mplus User's Guide. Seventh Edition. Los Angeles, CA: Muthén & Muthén.

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory (3rd edition). New York: McGraw-Hill.

Perrone M (2006). Differential item functioning and item bias: Critical consideration in test fairness. Applied Linguistics, 6(2): 1-3.

Hambleton, R., & Rodgers, J. (1995). Item bias review. Practical Assessment, Research, and Evaluation, 4(6). Retrieved at http://PAREonline.net/getvn.asp?v=4andn=6.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Nielsen and Lydiche.

Rost, J., & von Davier, M. (1994). A conditional item-fit index for Rasch models. Applied Psychological Measurement, 18(2), 171–182.

Schulz, W. & Fraillon, J. (2011). The Analysis of Measurement Equivalence in International Studies using the Rasch Model. Educational Research and Evaluation, 17(6), 447-464.

Schulz, W. (2009). Questionnaire Construct Validation in the International Civic and Citizenship Education Study. In: IERI Monograph Series Volume 2, 113-135.

Schulz, W., Ainley, J., Fraillon, J., Kerr, D. & Losito, B. (2010). ICCS 2009 International Report. Civic knowledge, attitudes and engagement among lower secondary school students in thirty-eight countries. Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

Schulz, W., Ainley, J., Fraillon, J., Losito, B. & Agrusti, G. (2016). IEA International Civic and Citizenship Education Study 2016. Assessment Framework. Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

Schulz, W., Fraillon, J., Ainley, J., Losito, B., & Kerr, D. (2008). International Civic and Citizenship Education Study assessment framework. Amsterdam: International Association for the Evaluation of Educational Achievement (IEA).

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. British Journal of Mathematical and Statistical Psychology, 27(2), 229–239.

van de gaer, E., Grisay, A., Schulz, W., & Gebhardt, E. (2012). The Reference Group Effect: An Explanation of the Paradoxical Relationship between Academic Achievement and Self-confidence across Countries. Journal of Cross-Cultural Psychology, 43, 1205-1228.

Van de Vijver, F. J. R., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An overview. European Review of Applied Psychology, 47, 263–279.

Wright, B.D., & Masters, G.N. (1982). Rating Scale Analysis: Rasch Measurement, Chicago.

Wu, M.L., Adams, R.J., Wilson, M.R., & Haldane, S. (2007). ACER ConQuest: General Item Response Modelling Software [computer program]. Camberwell, Vic.: Australian Council for Educational Research.