Questionnaire construct validation in the International Civic and Citizenship Education Study

Wolfram Schulz
Australian Council for Educational Research, Camberwell, Victoria, Australia

International studies tend to use student, teacher, and/or school questionnaires to collect contextual data on student and teacher characteristics, background, and activities, and the school's learning environment. Student measures of values, attitudes, and behavioral intentions are also often viewed as important learning outcomes, in particular within the context of studies of civic and citizenship education. The scaling of questionnaire items to obtain measures of students', teachers', and principals' perceptions and attitudes should therefore ideally be subject to a thorough cross-country validation of the underlying constructs. However, whereas those conducting international studies spend considerable effort ensuring measurement equivalence for international test instruments, the issue of equivalency of questionnaire data does not always receive the same attention. Using a set of student questionnaire items as an example, this article describes different methodological approaches to assess cross-national construct validation. With reference to an example from the field trial analyses for the IEA International Civic and Citizenship Education Study (ICCS), the article also shows the extent to which classical item statistics, factor analysis, and item response modeling help to assess the construct validity of questionnaire data obtained from international studies.

INTRODUCTION

The International Civic and Citizenship Education Study (ICCS) is the third international IEA study designed to measure the context and outcomes of civic and citizenship education, and it is explicitly linked through common questions to the IEA Civic Education Study (CIVED), which was undertaken in 1999 and 2000 (Amadeo, Torney-Purta, Lehman, Husfeldt, & Nikolova, 2002; Torney-Purta, Lehmann, Oswald, & Schulz, 2001; Schulz & Sibberns, 2004). The study, which builds on CIVED and is being conducted during 2008 and 2009, surveys 13- to 14-year-old students in 38 countries and will report on student achievement and perceptions related to civic and citizenship education.[1] Outcome data are being obtained from representative samples of students in their eighth year of schooling; context data are being collected from the students, their schools, and their teachers as well as through the study's national centers.

The aim of ICCS is to gather data on (a) student knowledge, conceptual understanding, and competencies in civic and citizenship education; (b) student background characteristics and participation in active citizenship; and (c) student perceptions of aspects of civics and citizenship. Instruments used in ICCS include an online national context survey completed by the national centers, a student test, a student questionnaire, a teacher questionnaire, and a school questionnaire. The ICCS assessment framework (Schulz, Fraillon, Ainley, Losito, & Kerr, 2008) outlines the aspects addressed in the cognitive test and the student perceptions questionnaire and provides a mapping of factors that might influence outcome variables and explain their variation.

It is recognized that there is substantial diversity in the field of civic and citizenship education within and across countries. Consequently, maximizing the involvement of researchers from participating countries in this international comparative study has been of particular importance for the success not only of this study in general but also for the process of developing an assessment framework and instruments. Input from national research centers has been sought throughout the study, and strategies were developed to maximize country contributions from the time of the early piloting activities through to selection of the final main survey instruments in June 2009. The main data collection was carried out between October and December 2008 in the Southern Hemisphere and between February and May 2009 in the Northern Hemisphere.

The students surveyed at these times were students enrolled in the grade that represents eight years of schooling, counting from the first year of primary school,

1 ICCS is managed by a consortium of three partner institutions (the Australian Council for Educational Research, the National Foundation for Educational Research in the United Kingdom, and the Laboratorio di Pedagogia Sperimentale at the Roma Tre University). The three institutions work in close co-operation with the IEA Secretariat, The IEA Data Processing and Research Center, and the national research coordinators (NRCs) of the participating countries. Further information about ICCS can be found at http://iccs.acer.edu.au/.

provided that their mean age at the time of testing was at least 13.5 years. According to this definition, for most countries the target grade was the eighth grade, or its national equivalent. The rationale for this definition was to have cross-nationally comparable student populations that are of the same age but also have similar levels of schooling.

One important feature of the ICCS data collection was measurement of value beliefs, attitudes, and behavioral intentions. This was typically done by administering questionnaires that included sets of four-point Likert-type items scaled to derive measures of latent constructs. Consequently, the comparability of these constructs became an important requirement for the ICCS data collection.

Language differences can have a powerful effect on equivalence (or non-equivalence). As with most international studies (see, for example, Chrostowski & Malak, 2004; Grisay, 2002), ICCS implements reviews of national adaptations and rigorous translation verifications to achieve a maximum of "linguistic equivalence." However, even slight deviations in wording (sometimes due to linguistic differences between source and target language) can lead to differences in item responses (see Harkness, Pennell, & Schoua-Glusberg, 2004; Mohler, Smith, & Harkness, 1998).

Non-equivalence in international studies can also be caused by cultural differences among the participating countries. Cultural habits can influence the degree to which respondents endorse certain item statements. In addition, differences between education systems (with different instructional practices and policies) and curricula can influence how respondents understand and interpret questionnaire items. For example, student responses indicating unfavorable learning conditions (such as disruptions at the beginning of each lesson) can be interpreted differently depending on what is commonly experienced in the national context (for an example, see Schulz, 2003).

According to van de Vijver and Tanzer (1997), while instruments might work properly, the cultural characteristics of groups of respondents can introduce bias in measurement. Byrne (2003) distinguishes three types of bias in cross-national research:

1. *Construct bias* refers to cases where a construct may be meaningful in one country, but not in another country;
2. *Method bias* refers to cases where data are biased by differences in responses to the instruments caused by cultural traits; and
3. *Item bias*, which refers to bias occurring at the level of the individual item. Constructs might be well measured in general, but some items may exhibit differential item functioning due to cultural differences.

Confirmatory factor analysis (CFA), which is based on the analyses of variances and covariances, provides an important tool for reviewing the cross-cultural validity of questionnaire constructs (see Kaplan, 2000). Little (1997) proposed extending the use of CFA to multiple-group analysis of mean and covariance structures (MACS) in order to test the comparability of measurement equivalence of psychological constructs and thereby detect possible socio-cultural variation of factor loadings and intercept parameters. Item response modeling (ref. item response theory or IRT; see Hambleton, Swaminathan, & Rogers, 1991) has also been used to detect non-equivalence of questionnaire constructs across countries (see, for example, Schulz, 2006; Wilson, 1994). IRT has furthermore been used to review the existence of different response patterns across cultural contexts in studies employing data collection instruments containing Likert-type items (Walker, 2007).

This article documents how both analyses of variances and covariances as well as item response modeling were used to address questionnaire construct validity and measurement equivalence during analysis of the ICCS field trial data. The analysis included reviews of item dimensionality, item/scale characteristics, and the measurement equivalence of model parameters across the participating countries.

DATA AND METHODS

The international field trial for ICCS was carried out in 32 participating countries between October 2007 and January 2008. On average, about 25 schools with about 600 students in the target grade in intact classrooms were selected. The following international instruments were used in the field trial:

- *The international student test:* This comprised 98 items in six different clusters administered in a completely rotated design with six randomly allocated booklets, each consisting of three 20-minute clusters;
- *The international student questionnaire:* Containing 71 background and 201 perceptions items, this was administered in three randomly allocated questionnaire forms;
- *The international teacher questionnaire:* This contained around 32 questions that took about 30 minutes to answer; and
- *The international school questionnaire:* This contained 22 questions that took 20 to 30 minutes to answer.

In addition, regional field trial instruments were administered in Europe and Latin America. These instruments consisted of short knowledge tests and questionnaire material designed to capture region-specific knowledge and perceptions.

The following verification procedures were implemented before the international field trial to ensure the highest possible level of instrument comparability:

- *Review of national adaptation:* During the first stage, the national centers submitted national adaptation forms (NAFs) for all instruments to the international study center (ISC) for review. ISC staff members reviewed the adaptations and sent the forms back with recommendations for further improvement, where appropriate. These forms were particularly useful as references during further instrument verification steps and data processing.
- *Translation verification:* After implementing suggestions from the adaptation review, the national centers submitted all instruments to be verified by professional language experts. The IEA Secretariat coordinated this activity. The verification outcomes were sent back to the national centers with (where appropriate) suggestions for improving the translations.

- *Layout verification:* After implementing suggestions from translation verification, the national centers assembled the final field trial instruments and submitted them for final layout verification by the ISC. The results of this final check were sent back to the countries.

The ICCS field trial analyses were based on a data collection in 718 schools in 31 countries and comprised questionnaire data from 19,369 students, 9,383 teachers, and 681 school principals. The analyses presented in this article focus on a set of items included in the student questionnaire and measuring students' expected civic participation in the future. This set of items is used to illustrate the following approaches, scope, and interpretation.

- *Exploratory factor analysis (EFA):* This was used at the preliminary analyses stage in order to review expected dimensionality of questionnaire items and to make preliminary decisions regarding the allocation to scales.
- *"Classical" item and scale statistics (such as reliabilities and item-total correlations):* These were computed to provide information on scaling characteristics and to permit a country-by-country review of item and scale performance.
- *Confirmatory factor analysis (CFA):* This was estimated for the pooled sample and separately for country sub-samples. These analyses were used not only to review the measurement model but also to review its fit as well as correlations between the latent variables across countries.
- *Multiple-group CFA:* This was estimated with different constraints to test measurement invariance more systematically across countries. However, multiple-group modeling was not systematically implemented in the analyses of ICCS field trial data.
- *Rasch modeling:* This provided information on item fit as well as estimates of item-by-country interaction with regard to the item location parameters.

This article discusses the results of these different analysis steps with regard to their usefulness for reviewing scale/item characteristics and validating constructs in international studies. To gain full benefit from this discussion, it is important to distinguish the following criteria:

- *Item performance:* Single items can be judged with regard to their appropriateness for measuring a construct both internationally and for individual education systems.
- *Construct measurement: S*cales can be scrutinized with regard to the extent to which it is possible to measure a certain construct reliably using a set of indicators.
- *Cross-country validation:* Constructs may be measured reliably in one country but not in another, which raises the question of whether, and to what extent, it is possible to measure the same construct with an internationally defined measurement model.

Each of the analysis approaches provides different pieces of information helpful to readers wanting to determine if items and scales used in an international study comply with their national criteria.

ICCS Field Trial Data Analysis

Exploratory Factor Analysis

In this article, the procedures for the ICCS questionnaire constructs are illustrated through analysis of a set of items measuring students' expectations about their future participation in political activities as an adult or as adolescent. Because of missing data, not all country datasets could be included in the analyses of these items. Table 1 shows the wording of the items used in the analyses. Expected participation in political life as an adult (Question I03) was measured with a set of seven items. Two dimensions were expected: expected electoral participation (scale name: VOTEPART, items I03a to I03c) and expected active political participation (scale name: POLPART, items I03d to I03g). Expected participation as an adolescent in the near future (Question I04) was measured with seven items expected to form a scale measuring expected informal civic participation (INFPART).

Table 1: Items measuring students' expected civic participation

Item	Question/item wording	Expected scale
Question I03	Listed below are different ways adults can take an active part in political life. **When you are an adult, what do you think you will do?**	
I03A	Vote in <local elections>	VOTEPART
I03B	Vote in <national elections>	VOTEPART
I03C	Get information about candidates before voting in an election	VOTEPART
I03D	Help a candidate or party during an election campaign	POLPART
I03E	Join a political party	POLPART
I03F	Join a trade union	POLPART
I03G	Stand as a candidate for a local or city office	POLPART
Question I04	Listed below are different actions that you as a young person could take during the next few years. **What do you expect that you will do?**	
I04A	Volunteer time to help people in the <local community>	INFPART
I04B	Collect money for a social cause	INFPART
I04C	Talk to others about your views on political and social issues	INFPART
I04D	Try to get friends to agree with your political opinions	INFPART
I04E	Write to a newspaper about political and social issues	INFPART
I04F	Contribute to an on-line discussion forum about social and political issues	INFPART
I04G	Join an organization for a political or social cause	INFPART

Note: Response categories were (1) I will certainly do this, (2) I will probably do this, (3) I will probably not do this, and (4) I will certainly *not* do this.

During the first stage of the ICCS field trial analysis, exploratory factor analyses were undertaken to review the expected dimensionality of questionnaire items following a review of item frequencies for valid and missing categories. The pooled international sample was used for these preliminary analyses, and the first decisions were made about the mapping of items to constructs for further analyses. Items were analyzed using principal component analyses with PROMAX rotation, which allows factors to be correlated. The software package MPLUS was used to estimate the results (Muthén & Muthén, 2001).[2]

Table 2 shows the results of the EFA for the items measuring students' expected political participation. The expected three-factor solution had an unsatisfactory model fit. The results for a four-factor solution clearly show that items I04A (volunteering time) and I04B (collecting money) loaded on a different factor than informal participation. As there were only two items measuring a fourth construct, "expected community

Table 2: EFA results for expected civic participation items (factor loadings for four-factor solution and factor correlations)

Item		Factors			
		1	2	3	4
I03A	Vote in <local elections>	**0.85**	0.01	- 0.03	0.01
I03B	Vote in <national elections>	**0.95**	- 0.06	- 0.02	- 0.05
I03C	Get information about candidates before voting	**0.56**	0.08	0.06	0.05
I03D	Help a candidate or party during campaign	0.16	0.06	0.09	**0.46**
I03E	Join a political party	- 0.03	- 0.05	- 0.02	**0.90**
I03F	Join a trade union	- 0.01	- 0.02	0.05	**0.72**
I03G	Stand as a candidate for a local or city office	- 0.04	0.04	0.07	**0.69**
I04A	Volunteer time to help people	- 0.03	**0.79**	0.00	0.02
I04B	Collect money for a social cause	- 0.03	**0.80**	0.01	- 0.06
I04C	Talk to others about your views	0.10	0.18	**0.52**	0.01
I04D	Try to get friends to agree with your opinions	0.01	0.04	**0.61**	0.04
I04E	Write to a newspaper	- 0.04	- 0.02	**0.81**	0.01
I04F	Contribute to an on-line discussion forum	- 0.01	- 0.10	**0.85**	- 0.04
I04G	Join an organization for a political or social cause	- 0.06	0.09	**0.61**	0.14
	Correlations between factors				
	Factor 1	**1.00**			
	Factor 2	**0.42**	**1.00**		
	Factor 3	**0.41**	**0.61**	**1.00**	
	Factor 4	**0.44**	**0.46**	**0.64**	**1.00**

Note: PROMAX rotation with maximum likelihood estimation based on pooled international field trial sample; RMSEA = 0.051, RMR = 0.018. Factor loadings > 0.4 in **bold**.

2 In general, maximum likelihood estimation was used for the majority of items with four categories in the ICCS field trial analyses. A mean- and variance-adjusted weighted least square (WLS) estimator was used for items with fewer categories.

participation," these two items were not viewed as sufficient for construct measurement and were discarded from the subsequent scaling analyses. The results also show quite strong positive correlations across the four factors. However, the size of these correlations still indicates that this item-set measures four clearly distinguishable factors.

Using EFA as a first analysis step is useful for confirming whether expectations about item dimensionality and scaling are reasonable. For the ICCS field trial analysis, EFA was used only with the pooled international sample because comparing results for each country dataset could have become complex. Also, CFA analysis was viewed as more appropriate for determining whether expectations regarding item dimensionality held for individual country data.

Classical Item and Scale Analysis

Once the preliminary analysis of dimensionality had been undertaken, the expected mapping of items to scales was revised according to the results of the exploratory factor analyses. Based on the revised item classification of scaled items, the following classical item statistics were computed for the pooled dataset and separately for each country.

- *Item-total correlations:* Pearson correlation coefficients between each item and the (corrected) overall raw score are a particularly useful means of reviewing translation errors. For example, a negative correlation with the overall score may indicate that a negatively phrased item ("Students of my age are too young to have a say in school matters") was translated as a positive one ("Students of my age have a say in school matters").
- *Scale reliabilities (Cronbach's alpha):* This coefficient gives an estimate of the internal consistency of each scale. For scales that are not used for individual test scores, but rather for group-level comparisons, we may refer to values over 0.7 as a satisfactory reliability and values over 0.8 as a high reliability. However, it is important to note that the coefficient is influenced by the number of items included in the scale.

Table 3 shows an example of classical item statistics for three items measuring students' expected participation in activities related to elections in each of the participating countries that had sufficient data. The table also shows the median statistics across the countries. In the table, each participating country has printed beside it the scale reliability (Cronbach's alpha), the number of items, the corrected item-score correlations, the number of cases, the percentage of missing responses, the mean of the raw scale (taking the average of all items), and the correlation of the raw score with the student performance in the test of civic knowledge.

Both the scale reliabilities and the item-total correlations indicated a high degree of consistency across countries. For all three items in most of the countries, the percentage of missing values did not exceed two. Only one country appeared to have a considerable proportion of students with no response. In most countries, there was a positive correlation between expected electoral participation and civic knowledge and understanding as measured by the international cognitive test.

Table 3: Classical item statistics for items measuring expected electoral participation (VOTEPART)

Country	Alpha coefficient	Items	Items: ISR103A	Items: ISRI03B	Items: ISRI03C	Valid numbers of cases	% missing responses	Scale mean	Correlation with test performance
CNT1	.727	3	.671	.654	.352	351	.85	1.82	.319
CNT2	.790	3	.687	.688	.529	418	1.65	2.10	.241
CNT3	.849	3	.736	.784	.639	339	2.87	2.17	.323
CNT4	.893	3	.800	.850	.724	482	1.43	2.00	.096
CNT5	.763	3	.611	.651	.524	516	3.37	2.40	.183
CNT6	.853	3	.776	.768	.641	158	2.47	2.18	.459
CNT7	.704	3	.576	.594	.409	301	22.62	2.14	-.016
CNT8	.861	3	.779	.798	.638	121	.82	1.91	.230
CNT9	.862	3	.776	.824	.625	406	1.69	2.11	.208
CNT10	.764	3	.652	.685	.466	335	.30	2.04	.281
CNT11	.804	3	.658	.726	.573	415	.48	2.16	.237
CNT12	.771	3	.685	.674	.479	402	.50	2.29	.390
CNT13	.779	3	.673	.689	.505	361	1.63	2.18	.410
CNT14	.792	3	.672	.679	.567	395	.75	2.38	.479
CNT15	.774	3	.719	.673	.461	351	.85	2.20	.120
CNT16	.817	3	.697	.712	.601	542	4.58	1.78	.279
CNT17	.827	3	.708	.719	.629	369	1.34	2.13	.291
CNT18	.755	3	.632	.648	.482	574	5.12	2.36	.263
CNT19	.725	3	.597	.676	.399	190	2.06	1.89	.318
CNT20	.870	3	.773	.796	.687	589	2.97	1.89	.357
CNT21	.879	3	.814	.866	.633	182	3.19	2.31	.340
CNT22	.858	3	.776	.798	.633	448	2.40	1.97	.397
CNT23	.788	3	.690	.730	.486	369	.27	2.15	.272
CNT24	.809	3	.695	.695	.590	352	5.88	2.37	.192
CNT25	.768	3	.644	.651	.529	380	1.55	2.26	.239
CNT26	.788	3	.677	.693	.526	417	.48	1.89	.370
CNT27	.842	3	.745	.780	.602	403	1.47	2.16	.337
CNT28	.874	3	.776	.829	.676	405	1.46	2.08	.486
CNT29	.849	3	.779	.769	.618	563	.71	2.41	*N/A*
CNT30	.873	3	.832	.793	.654	553	.90	2.05	.343
Median	**.806**	**3**	**.696**	**.715**	**.581**		**1.511**	**2.15**	**.291**

Note: Items were coded to values 0 (I will certainly *not* do this), 1 (I will probably *not* do this), 2 (I will probably do this), and 3 (I will certainly do this). N/A = not available.

For the scale derived from the items measuring expected political participation, VOTEPART showed good internal consistencies across participating countries, with a median Cronbach's alpha coefficient of 0.81. Similar results were obtained for POLPART and INFPART. The median reliabilities across countries were 0.81 and 0.83 respectively.

Tables featuring classical item statistics are useful for reviewing the performance of items and scales for each individual country dataset. The national centers of the participating countries were given guidelines on interpretation and were asked to review their own national item statistics. In cases of low reliabilities, inverted item-by-total correlations, and any other unusual findings for a particular country, the ISC asked the national center staff of the country concerned if translation problems or any specific aspects of the country's context could explain these deviations.[3]

Confirmatory Factor Analysis

The classical item tables used in ICCS are particularly useful for assessing the internal consistency of scales and the extent to which items in individual countries have similar correlations with the raw score. However, the review of classical item statistics does not allow a test of how well the expected model fits the observed data. To achieve this purpose, CFA can be carried out through the use of structural equation modeling (SEM) techniques (see Kaplan, 2000).

Within the SEM framework, latent variables are linked to observable variables via measurement equations. An observed variable x is modeled as

(1) $x = \Lambda_x \xi + \delta$,

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 variables. The expected covariance matrix is fitted according to the theoretical factor structure. With continuous variables, maximum likelihood (ML) estimation provides model estimates that try to minimize the differences between the expected (Σ) and the observed covariance matrix (S).

For CFA, an expected covariance matrix is fitted according to the theoretical factor structure. Model estimates can be obtained by minimizing the differences between the expected (*) and the observed covariance matrix (S). Measures for the overall fit of a model are then obtained by comparing the expected * matrix with the observed S matrix. If the differences between both matrices are close to zero, then the model "fits the data." If the differences are somewhat larger, the model "does not fit the data."[4]

3 Classical item statistics are usually also provided by most standard software (including ACER ConQuest) for item response modeling. SPSS macros were used to collate this information for the ICCS field trial analyses.

4 However, it needs to be noted that ML estimation assumes a normal distribution and continuous variables. Jöreskog and Sörbom (1993) therefore recommend for non-normal ordinal variables the use of WLS with polychoric correlation matrices and corresponding asymptotic covariance weight matrices. To simplify procedures for the ICCS field trial analyses, four-point Likert-type items were treated as continuous variables. This approach meant that standard software (such as the SAS CALIS procedure) could be used to estimate the results separately for each country in an efficient manner.

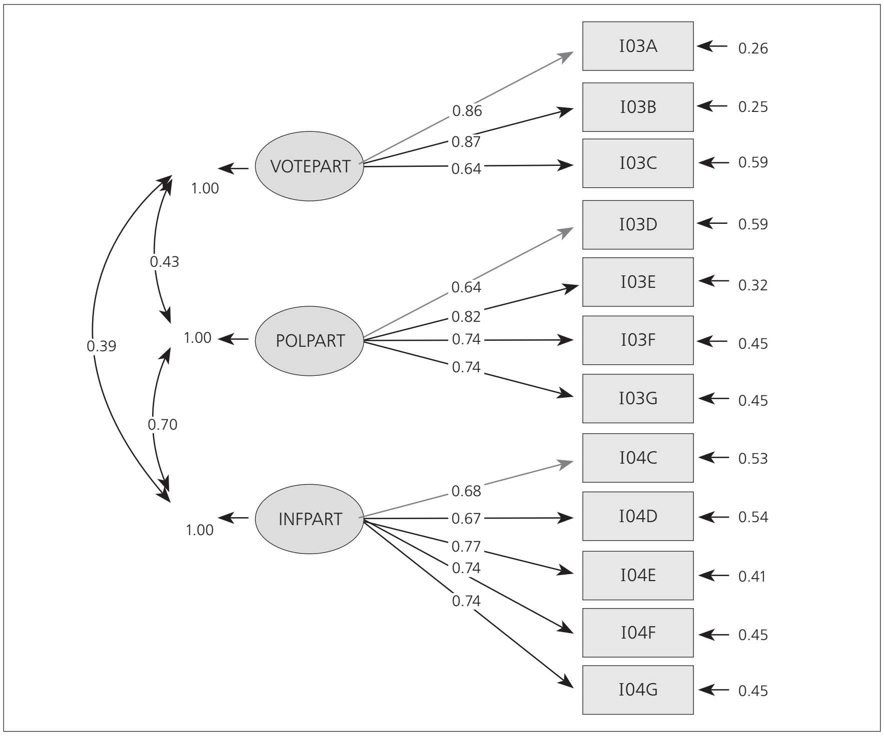
The following fit indices were used to assess the model fit of the CFA.

- The *root mean square error of approximation* (RMSEA): This measures the "discrepancy per degree of freedom for the model" (Browne & Cudeck, 1993, p. 144). A value of 0.05 and below indicates a close fit; values greater than 0.1 indicate a poor model fit.
- The *root mean square residual* (RMR): This has a similar interpretation, as the RMSEA with values below 0.05 indicate a close model fit.
- *Comparative fit index* (CFI) and the *non-normed fit index* (NNFI) (also known as the Tucker-Lewis Index, TLI): Like the RMSEA, these indices are less dependent than many other indices on sample size and are generally viewed as the most appropriate indices because they are less influenced by model complexity (see Bollen & Long, 1993). High values for CFI and TLI (over 0.9) indicate a satisfactory model fit.

Multi-dimensional CFAs were estimated in order to assess the relationships between the related latent dimensions measured in ICCS. CFA results for multi-dimensional models provide information about the appropriateness of the measurement models as well as about the correlation across the latent factors. A very high estimated latent correlation between two factors may indicate that the items measure reasonably similar dimensions and that a solution with fewer dimensions could be considered as an alternative model. If, for example, in a three-dimensional model, two of the dimensions are highly correlated, it might be more appropriate to assume a two-factor solution in which the items loading on the two highly correlated factors are used to measure only one combined factor.

Figure 1 shows a CFA for a three-factor model of ICCS items measuring students' expected participation. It shows a good model fit and positive correlations between the three latent dimensions. Although the correlation between POLPART and INFPART is very high (0.70), it indicates that both sets of items still reflect separate dimensions. Inspection of factor loadings reveals that some of the items measure the latent dimensions better than others: items I03C (VOTEPART) and I03D (POLPART) in particular have relatively lower factor loadings.

Figure 1: CFA results for items reflecting expected participation



Note: LISREL estimates with maximum likelihood estimation for pooled international sample. Items were coded to values 0 (I will certainly *not* do this), 1 (I will probably *not* do this), 2 (I will probably do this), and 3 (I will certainly do this). RMSEA = 0.070; RMR = 0.044; NNFI = 0.94; CFI = 0.95.

With international studies, it may not be appropriate to assume the same factor structure for each population. One way of looking at the consistency of factor structures is to use separate CFA for each country and to review model fit within each population in a comparative perspective. For the ICCS field trial analyses, separate models were estimated using the SAS CALIS procedure (see Hatcher, 1994). Table 4 presents the CFA results for expected participation items for each of the 30 country sub-samples. It also sets out the median values across the participating countries. The three-dimensional solution shows a satisfactory model fit in most countries, and the correlations between latent dimensions tend to be similar across countries.

Table 4: Comparison of model fit and latent correlation for items reflecting expected participation

	Model fit				Latent correlations		
					VOTEPART/	VOTEPART/	POLPART/
Country	RMSEA	RMR	CFI	NNFI	POLPART	INFPART	INFPART
CNT 1	0.12	0.08	0.88	0.88	0.33	0.28	0.63
CNT 2	0.08	0.06	0.94	0.94	0.42	0.41	0.65
CNT 3	0.10	0.05	0.90	0.90	0.45	0.45	0.60
CNT 4	0.08	0.06	0.95	0.95	0.49	0.39	0.68
CNT 5	0.06	0.04	0.96	0.96	0.46	0.23	0.69
CNT 6	0.08	0.05	0.93	0.93	0.45	0.41	0.74
CNT 7	0.07	0.07	0.95	0.95	0.61	0.49	0.81
CNT 8	0.10	0.06	0.92	0.93	0.50	0.52	0.63
CNT 9	0.08	0.06	0.95	0.95	0.48	0.43	0.78
CNT10	0.08	0.04	0.91	0.92	0.37	0.21	0.49
CNT11	0.08	0.03	0.94	0.94	0.41	0.32	0.53
CNT12	0.07	0.06	0.93	0.93	0.30	0.28	0.55
CNT13	0.10	0.07	0.92	0.92	0.34	0.51	0.73
CNT14	0.08	0.05	0.92	0.92	0.39	0.39	0.69
CNT15	0.08	0.06	0.93	0.93	0.29	0.24	0.64
CNT16	0.07	0.04	0.96	0.96	0.71	0.50	0.74
CNT17	0.06	0.04	0.95	0.95	0.40	0.30	0.69
CNT18	0.05	0.04	0.97	0.97	0.29	0.32	0.72
CNT19	0.09	0.07	0.90	0.90	0.28	0.21	0.64
CNT20	0.10	0.06	0.90	0.90	0.40	0.40	0.62
CNT21	0.11	0.07	0.93	0.93	0.23	0.29	0.80
CNT22	0.11	0.05	0.91	0.91	0.41	0.47	0.55
CNT23	0.09	0.07	0.91	0.91	0.41	0.25	0.57
CNT24	0.06	0.05	0.96	0.96	0.54	0.26	0.72
CNT25	0.08	0.05	0.94	0.94	0.35	0.38	0.79
CNT26	0.07	0.04	0.94	0.94	0.33	0.36	0.64
CNT27	0.09	0.06	0.91	0.92	0.30	0.36	0.53
CNT28	0.09	0.04	0.94	0.94	0.48	0.41	0.68
CNT29	0.09	0.04	0.93	0.93	0.23	0.42	0.64
CNT30	0.08	0.04	0.95	0.95	0.40	0.41	0.63
Median	**0.08**	**0.05**	**0.93**	**0.93**	**0.40**	**0.38**	**0.65**

Note: SAS CALIS estimates with maximum likelihood estimation. Items were coded to values 0 (I will certainly *not* do this), 1 (I will probably *not* do this), 2 (I will probably do this) and 3 (I will certainly do this).

Multiple-Group Analysis

A comparison of CFA results across countries shows the extent to which the measurement model fits the data for the pooled sample and in each dataset. In the case of multi-dimensional models, it also shows whether similar correlations between latent variables can be observed. However, it does not provide information about measurement invariance as such because country-specific models may fit the data but have different parameters.

To test parameter invariance, it is possible to use multiple-group modeling, which is an extension of standard SEM. If one considers a model where respondents belong to different groups indexed as g = 1, 2, ..., G, the multiple-group factor model becomes

(2) $x_g = \Lambda_{xg}\xi_g+\delta_g$,

A test of factorial invariance (H_Λ), where factor loadings are defined as being equal, can be written as

(3) $H_\Lambda : \Lambda_1 = \Lambda_1 = \Lambda_{2\ ...} = \Lambda_g$

Hypothesis testing using tests of significance tends to be problematic, in particular with data from large samples, where even smaller differences appear to be significant. Therefore, a modeling approach that looks at relative changes in model fit is preferable. This can be done by placing different equality constraints on parameters in multiple-group models and then comparing model-fit indices across different multiple-group models, each having an increasing degree of constraint, with the first having no constraints whatsoever. Different types of constraints can be used in order to review the invariance of model parameters. Once the invariance of factor structure and factor loadings has been confirmed, further constraints can be placed on the intercepts and factor covariances.

In the multiple-group analyses presented in this article, four different models were tested, with latent variables within each country having a mean of 0.[5] Because chi-square-based tests of statistical significance tend to be problematic with larger sample size, the results should be judged according to "relative model fit" of models with different degrees of constraints. The four models compared during the analysis of the ICCS field-trial data on expected civic participation were:

A. An unconstrained model with all parameters treated as country-specific;
B. A model with constrained factor loadings across countries;
C. A model with constraints on factor loadings and intercepts; and
D. A model with constraints on factor loadings, intercepts, and factor variances and covariances.

Table 5 shows the fit indices for the different multiple-group models for expected participation items RMSEA, NNFI, and CFI (for the overall model) as well as the median of RMR across individual country samples.

5 Because of the short timeline for analyzing the ICCS field-trial data, multiple-group analyses could not be fully implemented.

As is evident from the table, the differences in model fit between Model A (unconstrained) and Model B, with its constrained factor loadings, were minor, indicating that the assumption of constrained factor loadings is reasonable for the three-factor solution. With Model C, where the item intercepts were also constrained, the median RMR indicates that the indices fit is still satisfactory in a majority of countries but that the overall model fit is no longer satisfactory. The completely constrained Model D, where the factor variances and covariances were also assumed to be equal across countries, clearly does not fit the data. However, different factor variances and covariances can be viewed as a plausible finding, and it might therefore be unrealistic to expect that constructs related to expected participation have the same correlations regardless of the differences in political and civic culture between countries.

Table 5: Multiple-group models for expected participation items

	Model A	Model B	Model C	Model D
	Unconstrained	*Constrained loadings*	*Constrained loadings and intercepts*	*Completely constrained model*
RMSEA	0.08	0.08	**0.11**	**0.13**
NNFI	0.91	0.92	0.86	0.82
CFI	0.93	0.93	0.85	0.76
Median RMR across countries	0.05	0.06	0.07	**0.12**

Note: LISREL estimates with maximum likelihood estimation. RMR statistics indicate the fit for each individual country dataset; RMSEA, NNFI, and CFI indicate the fit for the overall model. RMR or RMSEA indices > 0.1 in bold.

Information from this type of analysis provides information about the extent of similarity for measurement models with different constraints across countries. When scaling data across countries, the assumption is made that items measure a given construct in the same, or at least in a very similar, way. Multiple-group modeling can be used to test this assumption in a more systematic way.

Item Response Modeling

Item response theory (IRT) (Hambleton et al., 1991) is often used in national or international large-scale studies for deriving individual scale scores for cognitive tests. Using IRT provides a large number of advantages for item analysis and allows the equating of different tests as well as the description of scales with test items. Increasingly, IRT is also used for deriving scale scores from questionnaire items (see Organisation for Economic Co-operation and Development, 2005; Schulz & Sibberns, 2004).

One example of how probabilities of responses to categorical items (e.g., Likert-type items) can be modeled is the *Partial Credit Model* (Masters & Wright, 1997), which is defined as

$$(4)\quad P_{x_i}(\theta) = \frac{\exp \sum_{j=0}^{x_i} (\theta_n - (\delta_i + \tau_{ij}))}{\sum_{h=0}^{m_i} \exp \sum_{j=0}^{h} (\theta_n - (\delta_i + \tau_{ij}))} \quad x_i = 0,1,2,\ldots,m_i\ ,$$

where $P_{x_i}(\theta)$ 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 for each step j.

One common measure of item fit is the weighted mean-square statistic (*infit*), a residual-based fit statistic. Weighted *infit* statistics can be computed for both item and step parameters; values close to 1 indicate good item fit. Values above 1 show an item discrimination that is lower than expected, given the model, whereas values below 1 indicate that the item discrimination is higher than expected. Reviewing this residual-based item fit indicates the extent to which each item fits the item response model. However, there are no clear rules for acceptable item fit, and it is generally recommended that analysts and researchers interpret residual-based statistics with caution (see Rost & von Davier, 1994).

Equation (5) below shows that the part of the model related to the item consists of the item parameter δ_i for item i and the step parameter τ_{ij} for step j of item t. In the context of a cross-national study, item parameters are assumed to be the same, similar to the assumption made in relation to a (one-dimensional) multiple-group model with constrained parameters. Tests of parameter invariance across countries can be reviewed by calibrating items separately within countries and then comparing model parameters and item fit. But it is also possible to estimate group effects directly by including further parameters in the scaling model. A partial credit model that includes estimates of item-by-country interactions can be described with this equation:

$$(5)\quad P_{x_i}(\theta) = \frac{\exp \sum_{j=0}^{x_i} (\theta_n (\delta_i - \eta_c + \lambda_{ic} + \tau_{ij}))}{\sum_{h=0}^{m_i} \exp \sum_{j=0}^{h} (\theta_n - (\delta_i - \eta_c + \lambda_{ic} + \tau_{ij}))} \quad x_i = 0,1,2,\ldots m_i$$

For the purpose of measuring parameter equivalence across a group of countries c, an additional parameter for country effects λ_{ic} is added to the (constrained) model. However, to obtain proper estimates, it is also necessary to include the overall country effect (η_c) in the model.[6] Both item-by-country interaction estimates (λ_{ic}) and overall country effects (η_c) are constrained to having a sum of 0.

6 The minus sign ensures that higher values of the country group effect parameters indicate higher levels of item endorsement in a country. An even less constrained model could go one step further by adding a country interaction and replacing the term τ_{ij} with an interaction between country and step parameters τ_{ijc}. This would allow an estimation of separate step parameters for each country. However, reviewing and interpreting the results of such an analysis becomes rather difficult, which is why only the item-by-country interaction effect was analyzed.

Models with country interaction effects provide estimates of the degree of parameter invariance across countries or groups of countries. The degree of parameter variation across countries can be summarized to provide information about the degree of measurement equivalence. In the ICCS field trial analysis, the median of absolute values for item-by-country interaction effect was taken as an indicator of parameter invariance for each item. In addition, the minimum and maximum effects were displayed to demonstrate the range of deviations across countries.[7]

For the analysis of ICCS field trial data, the scaling software *ACER ConQuest* (Wu, Adams, Wilson, & Haldane, 2007) was used for parameter estimation. Table 6 presents the results of the IRT analyses for the three ICCS scales reflecting expected participation. The first and second columns show the scale and item names. The third and fourth columns show the overall (international) item parameter and its fit for the pooled field trial sample. The fifth column shows the median of absolute values of item-by-country estimates across countries, and the sixth and seventh columns show the minimum and maximum values of item-by-country estimates. Because the estimation of parameter estimates and fit indices is less reliable for short scales, the low number of items in each scale (three, four, and five in the VOTEPART, POLPART, and INFPART scales, respectively) has to be taken into account when interpreting these analyses.

Table 6: IRT results for items reflecting expected participation

		International calibration results		Item-by-country interaction (λ_{ic})		
Scale	*Item*	*Parameter* (δ_i)	*Item fit (infit)*	*Median of absolute values*	*Minimum value across countries*	*Maximum value across countries*
VOTEPART	I03A	- 0.183	0.92	0.16	- 0.600	0.562
VOTEPART	I03B	- 0.073	0.89	0.17	- 0.674	0.341
VOTEPART	I03C	0.257	1.21	0.19	- 0.719	0.840
POLPART	I03D	- 0.503	1.18	0.23	- 0.673	0.698
POLPART	I03E	0.235	0.86	0.14	- 0.581	0.398
POLPART	I03F	0.137	0.99	0.16	-1.050	0.521
POLPART	I03G	0.131	1.00	0.17	- 0.433	0.497
INFPART	I04C	- 0.601	1.05	0.16	- 0.581	0.478
INFPART	I04D	- 0.179	1.08	0.16	- 0.494	0.991
INFPART	I04E	0.280	0.93	0.12	- 0.356	0.321
INFPART	I04F	0.234	0.99	0.15	- 0.345	0.324
INFPART	I04G	0.265	1.02	0.19	- 0.778	0.470

Note: ACER ConQuest estimates. Items were coded to values 0 ("I will certainly *not* do this"), 1 ("I will probably *not* do this"), 2 ("I will probably do this"), and 3 ("I will certainly do this").

7 More detailed lists of effects by country were included in appendices to the field trial analysis report sent to the participating countries.

Overall, Table 6 shows that the items in the scales appear to fit well. Only Item I03C ("Get information about candidates before voting") in the VOTEPART scale and Item I03D ("Help a candidate or party during campaign") in the POLPART scale show some evidence of less than ideal fit. These results correspond to the findings that these items also have relatively lower item-score correlations (see Table 3) and factor loadings in the CFA (see Figure 1).

When looking at the individual item-by-country interactions, we see that only Item I03D stands out as having slightly higher median cross-country variation of item location parameters when compared to other items. The minimum and maximum values show that, for most items, there are (at least some) countries whose estimated (national) parameters deviate considerably (more than 0.3 logits) from the international ones.

As with multiple-group model analyses, IRT modeling also allows us to compare overall model fit with and without taking the item-by-country interactions into account. The differences between deviances (a statistic indicating how well the item response model fits the data) are displayed in Table 7. The comparison shows that, for all three comparisons of (uni-dimensional) scaling models with and without item-by-country interaction effects, there is a statistically significant difference, which means that the model with country-specific parameters fits better.[8] Furthermore, the overall improvement of model fit is relatively small for each of the three model comparisons (about 2% less deviance).

Table 7: Deviance statistics for IRT models with and without item-by-country interaction effects

	VOTEPART	POLPART	INFPART
Model with interaction	72930 (100)	107227 (137)	133252 (171)
Model without interaction	74656 (10)	109415 (13)	135659 (16)
Difference in deviance	1726	2188	2407
Difference in degrees of freedom	90	124	155
χ^2	0.000	0.000	0.000

Note: ACER ConQuest estimates.

Estimating the item-by-country interaction of item location parameters provides researchers with an indication of the extent to which they can reasonably assume that student responses can be scaled with international item parameters. Given the relatively large sample sizes used in international studies, tests of the statistical significance of item-by-country estimates would inevitably lead to rejection of the hypothesis of measurement invariance. Another aspect to consider is that there are no clear criteria regarding the amount of item-by-country effects that is still tolerable. However, inclusion of this information during screening of the field trial material

8 The difference in deviance follows a chi-square distribution in which the degrees of freedom are equal to the difference between the parameters used in both models.

means that item selection preference can be given to those items and scales that show less item-by-country interaction.

CONCLUSION AND IMPLICATIONS

The analyses presented in this article drew on ICCS field trial data to show how different analysis methods can be used for construct validation purposes in international studies. Combining different approaches such as "classical" item analyses, factor analyses, and item response modeling has the potential to provide a comprehensive means of reviewing the extent to which one may assume measurement equivalence for questionnaire constructs in international studies.

In the case of the ICCS items regarding students' expected civic participation in the future, it was possible to derive three reliable scales measuring *expected electoral participation, expected active political participation, and expected informal civic participation*, with all reflecting highly similar dimensions across participating countries. Exploratory factor analysis showed two items clearly related to a fourth dimension, which was not included in subsequent analysis. Confirmatory factor analysis (CFA) illustrated that the three latent dimensions measured with the remaining 12 items, although positively correlated, reflected three distinct constructs. And while the three-dimensional model was shown as appropriate across the participating countries, the results from the multiple-group models and the item response modeling of item-by-country interaction effects provided evidence for a noticeable amount of between-country variation.

The results of each of the analysis steps illustrated in this article are complementary, and many of the findings about item or scale performance give similar information. For example, item-total correlations, factor loadings, and IRT item fit tend to coincide with the extent to which items measure a given construct. Likewise, there are numerous indices of scale reliability: within the context of classical item analysis, Cronbach's alpha is probably the most widely used index. But indices of scale reliability could also be derived from factor loadings in a CFA, and there are different IRT-based measures of scale reliabilities.

The results of the exploratory and confirmatory factor analyses also tended to be highly similar with regard to the analysis of item dimensionality. However, although exploratory factor analysis (EFA) already gives sufficient information about the dimensionality of items, CFA provides a way of modeling data under the assumption of items loading on only one factor, thereby allowing analysts to correctly estimate the correlation between latent factors.

Item response models can be conceptualized and are mathematically equivalent to logistic confirmatory factor analyses (see Glöckner-Rist & Hoijtink, 2003). But whereas estimating CFA for questionnaire items assesses primarily the overall fit of the expected dimensional model for sets of items, using item response modeling focuses on the performance of individual items under a logistic item response model.

Both multiple-group models and IRT models of item-by-country effects provide information about the differences between international and/or country-specific item parameters. They also give an indication of the extent to which one can reasonably impose a "one size fits all" approach to the measurement model. In both cases, stringent tests of measurement invariance are difficult to implement, and there are no clear rules about "tolerable" levels regarding the lack of measurement invariance. Researchers accordingly have to rely on "rules of thumb" and reviews of comparative rather than hypothesis testing.

The sequence of analysis of the ICCS field trial data allowed the collection of useful information at each step, with each step serving a different purpose. At the initial stage, the EFA results gave preliminary results regarding the expected item dimensionality that informed the allocation of items to scales for subsequent analyses. Tables with classical item statistics then provided information about how scales and individual items worked in each of the participating countries and enabled the international study center (ISC) to flag problematic items at both the international and the country level. Classical item statistics had the advantage of being easily understood by national center staff.

The use of multi-dimensional CFA for the pooled international sample made it possible to construct a general description of the general measurement model for a given set of items and procurement of estimates of the correlations between latent variables. Estimating these models separately for each country set enabled assessment of the appropriateness of the dimensional model in comparative terms. Using multiple-group model CFA can be seen as an alternative to comparing separate CFA with a more advanced tool that provides a test of measurement invariance for models with different parameter constraints. Because of its greater complexity and general timeline constraints, this step was not systematically implemented in the ICCS field trial analysis.

Item response modeling gives a different perspective because it focuses on the appropriateness of modeling item responses via use of a logistic model instead of analyses of covariances. Given that items in many international studies are scaled using IRT, it is important to assess the appropriateness of the assumption of using the scaling model with internationally determined item parameters across countries.

Using a stepwise approach made it possible to achieve a high level of scrutiny with regard to the validity of measuring constructs with questionnaire items across countries. Each step provided some additional information regarding the performance of individual items, the scalability of items, and the measurement of constructs and their cross-country validity. The combination of these different sources of information provided a good basis for item selection for the ICCS main survey.

It needs to be acknowledged that some aspects regarding cross-national comparability raised in the literature could not be addressed within the scope of the ICCS field trial analyses presented in this article. For example, concerns exist regarding the general viability of using Likert-type items for construct measurement in cross-cultural studies because of differences in response patterns across countries (see, for example, Heine, Lehman, Peng, & Greenholtz, 2002).

As already observed in other international studies, the analysis of ICCS field trial data generally shows a noticeable extent of parameter variance across countries. Indeed, it would be somewhat ingenuous to assume that questionnaire items translated into different languages and administered in different cultures and education systems could ever be responded to in exactly the same way. The crucial question, however, is at what level parameter variation really becomes a problem and leads to biased results in comparative studies. Answers to questions such as this require further methodological research (using, for example, simulation studies) directed at comparing the impact of different levels of construct measurement equivalence on analysis results.

References

Amadeo, J., Torney-Purta, J., Lehmann, R., Husfeldt, V., & Nikolova, R. (2002). *Civic knowledge and engagement: An IEA study of upper secondary students in sixteen countries*. Amsterdam: International Association for the Evaluation of Educational Achievement (IEA).

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

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & S. J. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Byrne, B. M. (2003). Testing for equivalent self-concept measurement across culture. In H. W. Marsh, R. G. Craven, & D. M. McInerney (Eds.), *International advances in self-research: Speaking to the future* (pp. 291–314). Greenwich: Information Age Publishing.

Chrostowski, S. J., & Malak, B. (2004). Translation and cultural adaptation of the TIMSS 2003 instruments. In M. O. Martin, I. V. S. Mullis, & S. J. Chrostowski (Eds.), *TIMSS 2003 technical report* (pp. 93–108). Amsterdam: International Association for the Evaluation of Educational Achievement (IEA).

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.

Grisay, A. (2002). Translation and cultural appropriateness of the test and survey material. In R. J. Adams & M. Wu (Eds.), *PISA 2000 technical report* (pp. 57–70). Paris: OECD Publications.

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.

Hatcher, L. (1994). *A step by step approach to using the SAS system for factor analysis and structural equation modeling*. Cary, NC: SAS Institute.

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.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8 user's reference guide*. Chicago: Scientific Software International, Inc.

Kaplan, D. (2000). *Structural equation modeling: Foundation and extensions*. Thousand Oaks, CA: Sage.

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

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101–122). New York: Springer.

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. (2001). *Mplus: Statistical analysis with latent variables*. Los Angeles: Author.

Organisation for Economic Co-operation and Development (OECD). (2005). *Technical report for the OECD Programme for International Student Assessment*. Paris: OECD Publications.

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

Schulz, W. (2003). *Validating questionnaire constructs in international studies: Two examples from PISA 2000*. Paper presented at the annual meeting of the American Educational Research Association (AERA), Chicago, April 21–25, 2003.

Schulz, W. (2006). *Testing parameter invariance for questionnaire indices using confirmatory factor analysis and item response theory*. Paper presented at the annual meeting of the American Educational Research Association (AERA), San Francisco, April 7–11, 2006.

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).

Schulz, W., & Sibberns H. (Eds.). (2004). *IEA Civic Education Study technical report*. Amsterdam: International Association for the Evaluation of Educational Achievement (IEA).

Torney-Purta, J., Lehmann, R., Oswald, H., & Schulz, W. (2001). *Citizenship and education in twenty-eight countries*. Amsterdam: International Association for the Evaluation of Educational Achievement (IEA).

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.

Walker, M. (2007). Ameliorating culturally based extreme item tendencies to attitude items. *Journal of Applied Measurement, 8*(3), 267–278.

Wilson, M. (1994). Comparing attitudes across different cultures: Two quantitative approaches to construct validity. In M. Wilson (Ed.), *Objective measurement II: Theory into practice* (pp. 271–292). Norwood, NJ: Ablex.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. (2007). ACER ConQuest 2.0: *General item response modelling software* [computer program manual]. Camberwell, VIC: ACER Press.