Questionnaire Construct Validation in the International Civic and Citizenship Education Study

Wolfram Schulz, Australian Council for Educational Research, Email: schulz@acer.edu.au

Abstract

International studies tend to use student, teacher or school questionnaires for the collection of contextual data on student or teacher characteristics, background, activities and the school's learning environment. Furthermore, student measures of values, attitudes and behavioural intentions are also frequently viewed as important learning outcomes, in particular in the context of studies of civic and citizenship education. Data obtained from these instruments become frequently important predictors of student performance or are treated as learning outcome variables of interest.

Therefore, the scaling of questionnaire items to obtain measures of students', teachers' and principals' perceptions and attitudes should ideally be subject of a thorough cross-country validation of the underlying constructs. However, whereas international studies use to spend considerable efforts on ensuring measurement equivalence for international test instruments, the issue of equivalency of questionnaire data does not always receive quite the same attention.

Using a set of student questionnaire items as an example, this paper describes how measurement equivalence was reviewed in the field trial analysis for the IEA International Civic and Citizenship Education Study (ICCS) using different methodological approaches including factor analysis and item response modelling.

Keywords: *civic education, item response theory, structural equation modeling, ICCS*

Introduction

ICCS is the third international IEA study designed to measure 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 (Torney-Purta, Lehmann, Oswald and Schulz, 2001; Amadeo et. al., 2004; Schulz and Sibberns, 2004). The study will survey 13-to-14-year old students in 38 countries in the years 2008 and 2009 and report on student achievement and perceptions related to civic and citizenship education. Outcome data will be obtained from representative samples of students in their eighth year of schooling and context data from the students, their schools and teachers as well as through national centres. The study builds on the previous IEA study of civic education (CIVED) undertaken in 1999. Information about ICCS can be found at http://iccs.acer.edu.au/.[1]

It is recognised that there is substantial diversity in the field of civic and citizenship education within and across countries. Consequently, maximising the involvement of researchers from participating countries in this international comparative study has been of particular importance for the success of this study in the process of developing an assessment framework and instruments. Input from national research centres has been sought throughout the study and strategies have been developed to maximise country contributions from early piloting activities until the selection of final main survey instruments in June 2009.

The students surveyed for ICCS are students enrolled in the grade that represents eight years of schooling, counting from the first year of ISCED Level 1, provided the mean age at the time of testing is at least 13.5 years. According to this definition, for most countries the target grade would be the eighth grade, or its national equivalent. The aim of the survey 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 on-line national context survey completed by national centres, a student test, a student questionnaire, a teacher questionnaire and a school questionnaire.

The ICCS assessment framework (Schulz, Fraillon and Ainley, 2008) outlines the aspects that

---

[1] The study 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 Rome Tre University) who work in close cooperation with the IEA Secretariat, the *IEA Data Processing and Research Center* and the National Research Coordinators (NRC) in the participating countries.

are addressed in the cognitive test and student perceptions questionnaire and provides a mapping of factors that might influence outcome variables and explain their variation. The main data collection will take place between October and December 2008 in the Southern Hemisphere and between February and May 2009 in the Northern Hemisphere.

One important feature of the ICCS data collection is the measurement of value beliefs, attitudes and behavioural intentions. This typically done by administering questionnaire including sets of four-point Likert-type items that are scaled to derive measures of latent constructs. Consequently, the comparability of these constructs becomes in an important requirement for the ICCS data collection.

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

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

According to van de Veijver and Tanzer (1997), instruments might work properly but characteristics of cultural groups of respondents may introduce bias in measurement. Byrne (2003) distinguishes three different kind of bias in cross-national research:

- *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.
- *Item bias* refers to bias that occurs 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 analysis of variances and covariances provides a tool for reviewing the cross-cultural validity of questionnaire constructs (see Kaplan, 2000). Little (1997) proposes to extend the use of CFA to multiple-group analysis of mean and covariance structures (MACS) for testing the comparability of measurement equivalence of psychological constructs and detecting possible socio-cultural variation of factor loadings and intercept parameters.

Item response modelling (Item Response Theory) (see Hambleton, Swaminathan and Rogers, 1991), has also been used to detect non-equivalence of questionnaire constructs across countries, in particular with regard to different response patterns when using Likert-type items (Walker, 2007). Research comparing both methodological approaches (see for example Wilson, 1994; Schulz, 1996) suggests that IRT provides a more rigorous test of parameter invariance across countries than covariance-based methods and, consequently, has a higher likelihood of leading to the rejection of the hypothesis of measurement equivalence.

Data and Methods

The results are based on field trial data analysis for ICCS. An 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 target grade corresponds to the eighth year of schooling provided that the minimum age of students is 13.5.

The following international instruments were used in the field trial:

- The international student test with 98 items in six different clusters administered in complete rotated design with six randomly allocated booklets, each consisting of three 20-minutes clusters.
- The international student questionnaire (with a total 71 background and 201 perceptions items) was administered in three randomly allocated questionnaire forms.
- The international teacher questionnaire contains around 32 questions that took about 30 minutes to answer.
- The international school questionnaire contains 22 questions which 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 prior to the international field trial to ensure a highest possible level of instrument comparability:

- **Review of national adaptation**: At the first stage, national centres submitted national adaptation forms (NAF) for all instruments to the International Study Centre (ISC) for a review. ISC staff members reviewed the adaptations and send 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, national centres submitted all instruments to be verified by professional language experts. The IEA Secretariat coordinated this activity and verification outcomes were sent back to national centres with possible suggestions for improvement of the translations.
- **Layout verification**: After implementing suggestions from translation verification national centres assembled the final field trial instruments and submitted them for a final layout verification by the International Study Centre. 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, 9383 teachers and 681 school principals.[2] The analyses presented in this paper will focus on two questions in the student questionnaire.

The following types of analysis to assess cross-country construct validity were available for the ICCS field trial analyses:

- *Exploratory Factor Analysis* (EFA) was used at the preliminary analyses stage to review expected dimensionality of questionnaire items
- *"Classical" item and scale statistics* (like reliabilities and item-total correlations) were computed to provide information on scaling characteristics
- *Confirmatory Factor Analysis* (CFA) for the pooled sample and separately for country sub-samples

[2] Two participating countries had submitted their field trial data later and their data could not be included in the international field trial analyses.

- *Multiple-group CFA* estimated with different constraints to test measurement invariance across countries
- *IRT scaling analysis* providing estimates of item-by-country interaction

Due to the short timeline for the analysis not all of these analysis steps could be implemented for all types of data. In particular, multiple-group analyses could not be carried out as standard part of the international analysis procedures but are included in this paper to illustrate it as an additional tool to assess construct validation in international studies.

ICCS Field Trial Data Analysis

Exploratory Factor Analysis

In this paper the procedures for the ICCS questionnaire constructs will be illustrated with the analysis of a set of items measuring students' expectations about their future participation as an adult or as adolescent. Due to missing data not all country data sets 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 core. 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 that were expected to form a scale measuring expected informal civic participation (INFPART).

Table 1 Items measuring students' expected civic participation

Expected Scale	Item	
VOTEPART	I03a	Vote in <local elections>
VOTEPART	I03b	Vote in <national elections>
VOTEPART	I03c	Get information about candidates before voting in an election
POLPART	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
INFPART	I04a	Volunteer time to help people in the <local community>
INFPART	I04b	Collect money for a social cause
INFPART	I04c	Talking 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 organisation for a political or social cause

* 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.

At the first stage of ICCS field trial analysis exploratory factor analyses were undertaken to review the expected dimensionality of questionnaire items together following a review of item frequencies for valid and missing categories. Generally, the pooled international sample was used for these preliminary analyses and at this stage first decisions were made about the mapping of items to constructs for further analyses.

Items were analysed using Principal Component Analyses with PROMAX rotation, which allows factors to be correlated. The software package MPLUS was used for estimating results (Muthén and Muthén, 2001).[3]

Table 2 shows the results of the EFA for the items measuring students' expected political participation. The expected three-factor solution had a unsatisfactory model fit and the results for a four-factor solution clearly show that items I04A (volunteering time) and I04B (collecting money) load on a different factor than informal participation. Therefore it was decided to remove these two items and assume a three-dimensional structure for the remaining 12 items.

[3] Generally, maximum likelihood estimation was used for the majority of items with four categories. For items with fewer categories a mean- and variance- adjusted WLS estimator (WLSMV) was used (see Muthén, du Toit, and Spisic, 1997).

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

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	Talking 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 organisation for a political or social cause	-0.06	0.09	**0.61**	0.14

* PROMAX rotation with Maximum Likelihood estimation based on pooled international field trial sample; RMSEA = 0.051, RMR = 0.018.

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. In the case of the items measuring expected civic participation, two items (I04A and I04B) were removed from the INFPART scale.

Based on the revised item classification of scaled items, the following classical item statistics are computed for the pooled dataset and separately for each country:

- *Item-total correlations*: Pearson's correlations between each item and the (corrected) overall raw score are particularly useful to review 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. Values over 0.7 indicate a satisfactory reliability, values over 0.8 an excellent reliability. However, it should be noted 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. For each participating country, 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 is printed.

Both scale reliabilities and item-total correlations indicate a high degree of consistency across countries. There are less than 2 percent of missing values for all three items in most of the countries, only in one country there appear to be a considerable proportion of students with no responses. In most countries there is a positive correlation between expected electoral participation in civic knowledge as measured by the international cognitive test.

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

Country	Alpha	Items	ISRI03A	ISRI03B	ISRI03C	Valid N	% miss	SCALE mean	COR_TEST
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**

* 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.

Table 4 shows the reliabilities for the three scales across countries. All three scales have good internal consistencies in all participating countries.

Table 4 Reliabilities for scales reflecting expected participation

COUNTRY	VOTEPART	POLPART	INFPART
CNT1	**.727**	**.794**	**.866**
CNT2	**.790**	**.805**	**.829**
CNT3	**.849**	**.779**	**.823**
CNT4	**.893**	**.848**	**.869**
CNT5	**.763**	**.847**	**.841**
CNT6	**.853**	.683	**.824**
CNT7	**.704**	**.808**	**.817**
CNT8	**.861**	**.844**	**.875**
CNT9	**.862**	**.817**	**.837**
CNT10	**.764**	**.786**	**.810**
CNT11	**.804**	**.744**	**.862**
CNT12	**.771**	**.707**	**.722**
CNT13	**.779**	**.822**	**.873**
CNT14	**.792**	**.817**	**.783**
CNT15	**.774**	**.840**	**.807**
CNT16	**.817**	**.815**	**.873**
CNT17	**.827**	**.757**	**.810**
CNT18	**.755**	**.846**	**.826**
CNT19	**.725**	**.762**	**.823**
CNT20	**.870**	**.796**	**.827**
CNT21	**.879**	**.835**	**.911**
CNT22	**.858**	**.825**	**.852**
CNT23	**.788**	**.815**	**.819**
CNT24	**.809**	**.806**	**.811**
CNT25	**.768**	**.821**	**.855**
CNT26	**.788**	**.790**	**.815**
CNT27	**.842**	**.796**	**.803**
CNT28	**.874**	**.816**	**.862**
CNT29	**.849**	**.856**	**.817**
CNT30	**.873**	**.807**	**.870**
Median	**.806**	**.808**	**.826**
Number of items	*3*	*5*	*5*

* Cronbach's Alpha coefficients. Coefficient > 0.70 in **bold**.

Confirmatory Factor Analysis

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

(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 variances.

The expected covariance matrix is fitted according to the theoretical factor structure. With continuous variables, Maximum Likelihood (ML) estimation provides model estimates trying to minimise the differences between the expected (Σ) and the observed covariance matrix (S).

However, it should be noted that Maximum Likelihood (ML) estimation both require normal distribution and continuous variables. Jöreskog and Sörbom (1993) recommend for non-normal, ordinal variables to use Weighted Least Square Estimation (WLS) with polychoric correlation matrices and corresponding asymptotic covariance weight matrices.[4] Comparisons between the two types of estimation (ML versus WLS) show that for CFA highly similar results are obtained. To simplify procedures for the ICCS field trial analyses four-point Likert-type items were treated as continuous variables. By doing this standard software (like the SAS CALIS procedure) could be used for estimating country-by-country results.

For CFA, an 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 * 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".

There are no clear criteria for judging satisfactory model fit for CFA: Chi-square test statistic for the null hypothesis of *=S become rather poor fit measures with larger sample sizes because even small differences between matrices appear as significant deviations.

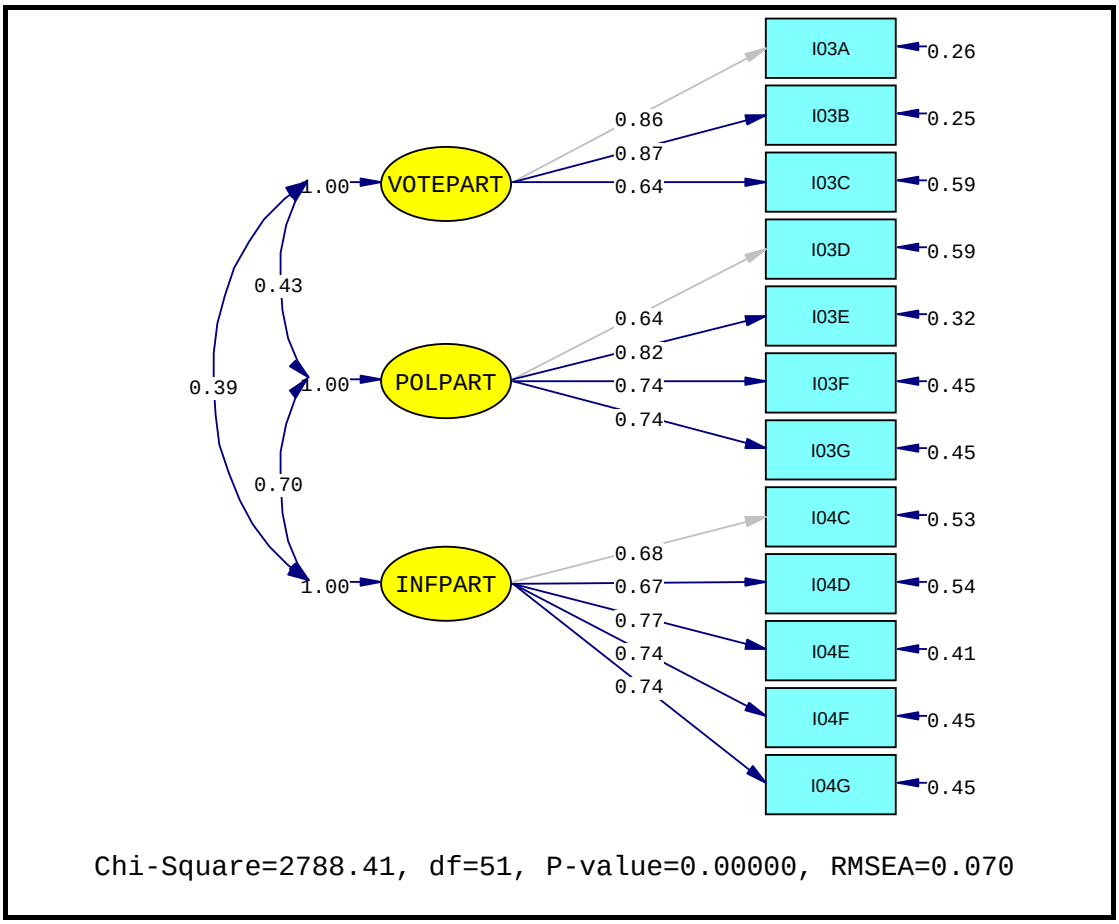
---

[4] Alternatively, mean-adjusted WLS estimator (WLSM) and mean- and variance- adjusted WLS estimator (WLSMV) that do not require large sample sizes are available in the Mplus software program (Muthén, du Toit, and Spisic, 1997).

Recent practice gives emphasis to alternative fit indices like the *Root Mean Square Error of Approximation* (RMSEA), which measures the "discrepancy per degree of freedom for the model" (Browne and Cudeck, 1993: 144). A value of .05 and less indicates a close fit and values greater than 1.0 indicate poor model fit. The *Root Mean Square Residual* (RMR) has a similar interpretation with RMR values of less than 0.05 indicating close model fit.

In addition, model fit can be assessed using the *Comparative Fit Index* (CFI) and the *Non-normed Fit Index* (NNFI) (also known as the Tucker-Lewis Index, TLI) which are less dependent on sample size and correct for model complexity (see Bollen and Long, 1993). High values for CFI and TLI (over 0.9) indicate satisfactory model fit.

Figure 1 CFA results for items reflecting expected participation



* 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).

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. The correlation between POLPART and INFPART is very high with 0.70 but it

indicates that both sets of items still reflect separate dimensions.

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. For the ICCS field trial analyses these models were estimated using the SAS CALIS procedure (see Hatcher, 1994).

Table 5 Comparison of model fit and latent correlation for items reflecting expected participation

	Model fit				Latent correlations		
COUNTRY	**RMSEA**	**RMR**	**CFI**	**NNFI**	**VOTEPART/ POLPART**	**VOTEPART/ INFPART**	**POLPART/ INFPART**
CNT1	0.12	0.08	0.88	0.88	0.33	0.28	0.63
CNT2	0.08	0.06	0.94	0.94	0.42	0.41	0.65
CNT3	0.10	0.05	0.90	0.90	0.45	0.45	0.60
CNT4	0.08	0.06	0.95	0.95	0.49	0.39	0.68
CNT5	0.06	0.04	0.96	0.96	0.46	0.23	0.69
CNT6	0.08	0.05	0.93	0.93	0.45	0.41	0.74
CNT7	0.07	0.07	0.95	0.95	0.61	0.49	0.81
CNT8	0.10	0.06	0.92	0.93	0.50	0.52	0.63
CNT9	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**

* SAS CALIS estimates with Maximum Likelihood estimation. RMSEA > 0.1 is **bold**. 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).

Table 5 shows the CFA results for expected participation items in the 30 country sub-samples

with sufficient data. The model fit is satisfactory in all but six countries and the correlations between latent dimensions are generally similar across countries.

Multiple-Group Analysis

To test parameter invariance, it is also possible to use multiple-group modelling, 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 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 form large samples where even smaller differences appear to be significant. Therefore, a modelling approach looking at relative changes in model fit is preferable. This can be done by setting placing different equality constraints on parameters in multiple-group models and comparing model fit indices across different multiple-group models with increasing constraints starting with a totally unconstrained model.

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 might be placed on intercepts and factor covariances.

In the multiple-group analyses presented in this paper four different models will be tested.[5] As 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.

Table 6 shows four different multiple-group models: Starting from a totally unconstrained model, in a first step factor loadings are constrained within groups of countries. In a second step factor loadings are constrained across all countries and in a third step additional constraints are placed on intercepts. The fourth model assumes also factor variances and

[5] Due to the short timeline multiple-group analyses could not be fully implemented in the field trial analysis for ICCS.

covariances to be equal across countries.[6]

Table 6 Description of multiple-group models in analysis

	Constraints
Model 1	Unconstrained model
Model 2	Constraints on factor loadings across countries
Model 3	Constraints on factor loadings and intercepts across countries
Model 4	Completely constrained model

Table 7 shows the results of the different multiple-group models for expected participation items. There are only minor differences in model fit between the unconstrained model and the model with constrained factor loadings. When constraining item intercepts, the fit indices still remain satisfactory in a majority of countries but the RMR values in a number of countries as well as the overall fit indices show an unsatisfactory model fit.

The completely constrained model where also factor variances and covariances are assumed to be equal across countries clearly does not fit the data. However, it appears that different factor variances and covariances are a quite plausible finding and not necessarily an indication of measurement invariance. It would be rather unrealistic to expect that constructs related to expected participation have the same correlations regardless of the differences in political and civic culture between countries.

In summary, the multiple-group analyses indicate that the dimensionality of the items measuring students' expected political participation is highly similar across countries. Even though the overall fit for the third model with constraints on intercepts is not longer satisfactory, the RMR in a majority of countries is not very different from the less constrained model.

[6] Further possible model variations could include constraining intercepts (thresholds in this case of using categorical items). However, similar response frequencies across countries were not viewed as a reasonable model assumption in an international study.

Table 7 Multiple-group Models for expected participation items

RMR for	Unconstrained	Constrained loadings	Constrained loadings and intercepts	Completely constrained model
CNT1	0.08	0.08	**0.12**	**0.14**
CNT2	0.06	0.06	0.06	0.07
CNT3	0.05	0.06	0.07	**0.12**
CNT4	0.06	0.07	0.07	**0.18**
CNT5	0.04	0.05	0.08	**0.14**
CNT6	0.05	0.06	0.08	**0.15**
CNT7	0.07	0.08	0.08	**0.34**
CNT8	0.06	0.07	0.07	**0.11**
CNT9	0.06	0.07	0.07	**0.13**
CNT10	0.04	0.04	0.04	**0.18**
CNT11	0.03	0.04	0.05	**0.19**
CNT12	0.06	0.08	**0.11**	**0.14**
CNT13	0.07	0.08	0.08	0.09
CNT14	0.05	0.06	0.08	0.10
CNT15	0.06	0.06	0.06	0.10
CNT16	0.04	0.05	0.05	**0.15**
CNT17	0.04	0.04	0.04	**0.12**
CNT18	0.04	0.05	0.05	**0.16**
CNT19	0.07	0.08	0.09	0.09
CNT20	0.06	0.07	0.06	**0.12**
CNT21	0.07	0.06	0.07	0.09
CNT22	0.05	0.06	0.06	**0.13**
CNT23	0.07	0.07	0.07	0.08
CNT24	0.05	0.06	**0.10**	**0.34**
CNT25	0.05	0.06	0.05	0.07
CNT26	0.04	0.05	0.05	0.07
CNT27	0.06	0.06	0.07	**0.11**
CNT28	0.04	0.05	0.05	**0.13**
CNT29	0.04	0.04	0.06	**0.11**
CNT30	0.04	0.05	0.05	0.07
Median	0.05	0.06	0.07	**0.12**
Overall fit				
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

* LISREL estimates with maximum likelihood estimation. RMR or RMSEA indices > 0.1 in **bold**.

Item Response Modelling

Probabilities of responses to categorical items (for example Likert-type items) can be modelled using the *Partial Credit Model* (Masters and Wright, 1997)[7], which is defined as

$$(5) \quad P_{x_i}(\theta) = \frac{\exp \sum_{k=0}^{x} (\theta_n - \delta_i + \tau_{ij})}{\sum_{h=0}^{m_i} \exp \sum_{k=0}^{k} (\theta_n - \delta_i + \tau_{ij})} \qquad x_i = 0,1,\ldots,m_i,$$

where $P_{xi}(\theta)$ 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. Item fit can be assessed using the weighted mean-square statistic (infit), a residual-based fit statistic. Weighted *infit* statistics can be computed both for item and step parameters.

IRT scaling methodology does not allow researchers to review the fit of scaling models for sets of items. Tests of parameter invariance across countries can be reviewed by calibrating items separately within countries and then comparing model parameters and item fit. In addition, it is possible to estimate group effects directly by including further parameters in the scaling model.

Equation (5) 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 i. When using the scaling software *ACER ConQuest* (Wu, Adams, Wilson and Haldane, 2007), the model term $\delta_i + \tau_{ij}$ is described with the statement ITEM+ITEM*STEP. For the purpose of the analysis of parameter equivalence, an additional parameters for country effects (CNT*ITEM) can be added to this model. However, to get proper estimates it is also necessary to include the overall country effect (CNT) in the model. Therefore, a Conquest model that estimates item-by-country interactions is defined as ITEM-CNT+ITEM*CNT+ITEM*STEP.[8]

---

[7] An alternative is the Rating Scale Model (RSM) which has the same step parameters for all items in a scale (see Andersen, 1997).

[8] 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 ITEM*STEP with an interaction between country and step parameters (CNT*ITEM*STEP). This would least to an estimation of separate step

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 summarised to inform about the degree of measurement equivalence. In the ICCS field trial analysis, the median 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.[9]

The IRT analyses for the three ICCS scales reflecting expected participation are shown in Table 8. The low number of items in each scale (three, four and five in the VOTEPART, POLPART and INFPART scales respectively) should be considered when interpreting these analyses.

Table 8 IRT results for items reflecting expected participation

		Calibration results		Item-by-country interaction		
Scale	Item	Parameter	Fit	Median of absolute values	Minimum	Maximum
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

* 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).

Overall the items in the scales appear to fit well. Only item I03C in the VOTEPART scale and item I03D in the POLPART scale show some evidence of less than ideal fit. Item I03D also has median item-by-country interaction DIF of 0.23 logits which together with the poor fit suggest that this is the weakest item in the POLPART scale.

---

parameters for each country. However, results of these analysis become rather difficult to review and therefore only the item-by-country interaction effect was analysed.

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

When looking at the item-by-country interactions, only Item D03D stands out as having a higher than ideal median item-by-country interaction DIF. However, it should be noted that generally item location parameters vary to a certain degree across countries.

Conclusion and Implications

This paper illustrates how construct validation was carried out for the ICCS field trial analysis. Combining "classical" item analyses, covariance-based analysis and item response modelling provides a comprehensive approach for reviewing the extent to which one may assume measurement equivalence for questionnaire constructs in international studies.

Researchers should be aware that using different methods will not provide the exactly the same results. However, item-by-total correlations and factor loadings in CFA tend to correspond to lack of item fit when applying item response models. Research has generally shown that IRT modelling appears to be more rigorous test of measurement equivalence than multiple-group modelling.

The example analysis in this paper correspond to findings from other research that there is usually some extent of parameter variance across countries. Indeed, it would be rather ingenuous to assume that questionnaire items translated into different languages and administered in different cultures and educational systems could be responded in exactly the same way. The crucial question is rather at what level parameter variation really becomes a problem and leads to biased results in comparative studies. Simulation studies may provide a way of exploring this issue by comparing the impact of different levels of construct measurement equivalence on analysis results.

International studies should preferably use questionnaire items that have lower levels of parameter variation across countries. The methods outlined in this paper provide tools to assess this matter at the field trial stage. In ICCS, indicators of measurement equivalence were one important criterion in the selection process for the final main survey instruments and assisted greatly the final revision of questionnaire material.

After the final data collection, further steps will be undertaken to investigate the issue of construct validation. Although at this stage results can no longer assist with the improvement of questionnaire material, they are still informative with regard to the interpretation of findings in the final data analysis and will be documented in the technical report for this study.

References

Amadeo, J., Torney-Purta, J., Lehmann, R., Husfeldt, V., and 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. and Long, S. J. (1993) (Eds.). *Testing Structural Equation Models*, Newbury Park/London.

Browne, M.W., and Cudeck, R. (1993). Alternative Ways of Assessing Model Fit. In: K.A. Bollen and S.J. Long (Eds.). *Testing Structural Equation Models*, Newbury Park/London, 136-162.

Byrne, B. M. (2003). Testing for Equivalent Self-Concept Measurement across Culture. In: H. W. Marsh, R. G. Craven and D. M. McInerney (eds.). *International advances in self-research: speaking to the future* (pp. 291-314). Greenwich: Information Age Publishing.

Chrostowski, S. J. and Malak, B. (2004). Translation and Cultural Adaptation of the TIMSS 2003 Instruments. In: Martin, M. O., Mullis, I. V. S. and Chrostowski, S. J. (eds). *TIMSS 2003. Technical Report* (pp. 93-108). Amsterdam: IEA.

Grisay, A. (2002). Translation and Cultural Appropriateness of the Test and Survey Material. In: Adams, R. and Wu, M. (eds.). *PISA 2000. Technical Report* (pp. 57-70). Paris: OECD Publications.

Hambleton, R. K., Swaminathan, H. and Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, London, New Delhi: SAGE Publications.

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

Harkness, J., Pennell, B., Schoua-Glusberg, A. (2004) Survey Questionnaire Translation and Assessment. In: Presser, S., Rothgeb, J., Couper, M., Lessler, J., Martin, E. and Singer, E. (eds.) *Questionnaire Development Evaluation and Testing Methods*. Hoboken: Wiley.

Jöreskog, K.G. and Dag Sörbom (1993). *LISREL 8 User's Reference Guide*. Chicago: SSI.

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

Little, T. D. (1997). Mean and Covariances Structures (MACS) Analyses of Cross-Cultural Data: Practical and Theoretical Issues. In: *Multivariate Behavioural Research*, 32 (1), 53-76.

Masters, G. N. and 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.

Mohler, P. P., Smith, T. W. and Harkness, J. A. (1998). Respondent's Ratings of Expressions from Response Scales: A Two-Country, Two-Language Investigation on Equivalence and Translation. In: Harkness, J. A. (ed.) *Nachrichten Spezial*, Cross-Cultural Survey Equivalence 3 (1998). Mannheim: ZUMA, 1998.

Muthén, L. K., and Muthén, B. O. (2001). *Mplus: Statistical analysis with latent variables*. Los Angeles: Muthén and Muthén.

Muthen, B. O., du Toit, S. H. C., & Spisic, D. (1997). *Robust inference using weighted least squares and*

quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Unpublished manuscript.

Schulz, W. (2003). *Validating Questionnaire Constructs in International Studies. Two Examples from PISA 2000*. Paper presented at the Annual Meetings of the American Educational Research Association (AERA) in Chicago, 21-25 April.

Schulz, W. (2006). *Testing Parameter Invariance for Questionnaire Indices using Confirmatory Factor Analysis and Item Response Theory*. Paper presented at the Annual Meetings of the American Educational Research Association (AERA) in San Francisco, 7-11 April.

Schulz, W., Fraillon, J. and Ainley, J. (2008) *International Civic and Citizenship Education Study. Assessment Framework*. Amsterdam: International Association for the Evaluation of Educational Achievement (IEA) (forthcoming).

Schulz, W. and Sibberns H. (eds.) (2004). *IEA Civic Education Study. Technical Report*. Amsterdam: 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. and 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 Measuremen*t 8(3) 2007, 267-278.

Wilson, M. (1994). Comparing Attitude 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. and Haldane, S. (2007). *ACER ConQuest 2.0: General item response modelling software* [computer program manual]. Camberwell, Vic.: ACER Press.