Educational Research and Evaluation: An International Journal on Theory and Practice

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/nere20

The analysis of measurement equivalence in international studies using the Rasch model

Wolfram Schulz [a] & Julian Fraillon [a]

[a] Australian Council for Educational Research , Melbourne, Australia

Published online: 30 Nov 2011.

To cite this article: Wolfram Schulz & Julian Fraillon (2011) The analysis of measurement equivalence in international studies using the Rasch model, Educational Research and Evaluation: An International Journal on Theory and Practice, 17:6, 447-464, DOI: 10.1080/13803611.2011.630559

To link to this article: http://dx.doi.org/10.1080/13803611.2011.630559

Educational Research and Evaluation
Vol. 17, No. 6, December 2011, 447–464

Routledge
Taylor & Francis Group

The analysis of measurement equivalence in international studies using the Rasch model

Wolfram Schulz* and Julian Fraillon

Australian Council for Educational Research, Melbourne, Australia

When comparing data derived from tests or questionnaires in cross-national studies, researchers commonly assume measurement invariance in their underlying scaling models. However, different cultural contexts, languages, and curricula can have powerful effects on how students respond in different countries. This article illustrates how the application of the Rasch item response theory (IRT) model (Rasch, 1960) can be used for assessing differences in measurement properties of tests and questionnaires with reference to examples from the field trial analyses for the International Association for the Evaluation of Educational Achievement (IEA) International Civic and Citizenship Education Study (ICCS). It also discusses the general scope and limitations of the analyses undertaken in the context of this study.

Keywords: measurement equivalence; Rasch modeling; international studies; civic and citizenship education; cross-cultural research

Introduction

Most comparative international studies on educational achievement make use of item response theory (IRT) models for the scaling of student responses to test items. Using IRT supports researchers to use rotated test booklets, equate tests, describe performance scales, and obtain proficiency estimates from multiple imputation (plausible values). In addition, an increasing number of studies employ IRT to obtain scale scores from questionnaire items. One important assumption for using IRT scaling (or other parametric scaling methods) in cross-national research is measurement invariance so that common parameters can be used to make comparisons across countries. Measurement invariance holds if individuals with the same score on the same measurement instrument have the same standing on the underlying construct that is measured. In comparative international studies, the source measurement instruments are the same for all countries, but each collects data using adapted and translated versions of the source instruments.

It is widely recognised that language differences may have powerful effects on measurement equivalence of test and questionnaire items, and this consequently challenges assumptions made about measurement invariance. Most international studies (see, e.g., Chrostowski & Malak, 2004; Grisay, 2002) implement rigorous

*Corresponding author. Email: schulz@acer.edu.au

ISSN 1380-3611 print/ISSN 1744-4187 online

http://dx.doi.org/10.1080/13803611.2011.630559
http://www.tandfonline.com

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. Furthermore, non-equivalence can also be caused by the cultural differences among participating countries in international studies (Byrne, 2003; Van de Veijver & Tanzer, 1997).

Whilst IRT modelling has become widely used to review measurement equivalence of test instruments in international studies (see, e.g., Olsen, Martin, & Mullis, 2008; Organisation for Economic Co-operation and Development [OECD], 2009), the measurement equivalence of questionnaire instruments has traditionally been investigated using confirmatory factor analysis (CFA) methods. Recently, it has become more common to also apply IRT methods (see Raju, Laffitte, & Byrne, 2002; Schulz, 2009; Walker, 2007).

In this article, we outline ways of obtaining estimates of measurement invariance applying the Rasch model and discuss their interpretations as well as limitations of this approach. The purpose of this article is to illustrate the potential of Rasch IRT modelling for the analysis of item-by-country interactions for test and questionnaire items using example field trial data from the International Association for the Evaluation of Educational Achievement (IEA) International Civic and Citizenship Education Study (ICCS).

Theoretical framework

For the analysis of ICCS data, IRT methods (see Hambleton, Swaminathan, & Rogers, 1991) were used as a general framework for the scaling of test and questionnaire items. In particular, the one-parameter (Rasch) model (Rasch, 1960) was applied as a model that predicts the probability of selecting the correct response of a test item depending on a latent trait θ.

For items with two categories (e.g., scored 1 for correct responses and 0 for incorrect responses), item responses are modelled as

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

where $P_i(\theta)$ is the probability of person n to score 1 on item i. θ_{n} is the estimated latent trait of person n and δ_{i} the estimated location of item i on this dimension. For each item, item responses are modelled as a function of the latent trait θ.[1]

In the case of polytomous items with more than two (k) categories this model can be generalised to the so-called partial credit model (Masters & Wright, 1997) as

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

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.

The goodness of fit for individual items can be determined by calculating a *mean square statistic* (Wright & Masters, 1982). Reviewing this residual-based item fit indicates the extent to which each item fits the item response model: A value of 1 indicates the best possible item fit to the Rasch model, values above 1 show an item discrimination which is lower, and values above 1 an item discrimination which is higher than expected. 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 and in conjunction with other indicators of item fit (see Rost & von Davier, 1994).

In view of the methodological shortcomings of IRT item fit statistics, in our review of the ICCS item scaling properties we used the *weighted mean square statistic* as one of several indicators. Items with weighted mean square statistics below 0.8 or above 1.2 were flagged, but the assessment of item fit was always undertaken in conjunction with other indicators of scaling properties. Additional indicators included averages of latent trait estimates within item categories or graphical displays (item characteristic curves)[2], as well as classical item statistics such as item-total correlations.

When applying a parametric measurement model, we generally assume parameter invariance across subgroups in the sample. Within the context of cross-national studies, we assume that the measurement model neither varies across subgroups within countries nor across countries. If this assumption does not hold, it becomes questionable whether measures across subgroups (here: national samples) are still comparable, and it would be uncertain to what extent any differences between them reflect variation in the latent trait or variation in measurement properties and potentially invalidate any conclusions drawn from cross-national comparisons of the data.

Measurement equivalence holds when a set of observed variables X, after controlling for a particular latent trait T, which they are designed to measure, is independent of any other variables V that could possibly cause measurement bias. It should be noted that any lack of measurement equivalence could also be viewed as an issue of multidimensionality where item responses depend not only on the latent trait but also on other factors (see Jak, Oort, & Dolan, 2010).

Within the framework of IRT, the issue of measurement variance is referred to as *differential item functioning* (DIF), which consists of finding different item parameters within different subgroups of a sample (see Hambleton & Rodgers, 1995; Perrone, 2006). For example, gender DIF occurs when a test item is relatively easier for girls than for boys or vice versa. Likewise, cross-national DIF or *item-by-country interaction* occurs when students from different countries but with the same ability vary in their probability of giving the correct answer to a test question.

Tests of parameter invariance across national subsamples can be reviewed by calibrating items separately within countries and then comparing model parameters and item fit across countries. Alternatively, it is also possible to estimate group effects directly by including further parameters as facets in the scaling model. For example, a partial credit model that includes estimates of item-by-country interactions can be described with the following equation:

$$P_{x_i}(\theta) = \frac{\exp \sum_{j=0}^{x_i} (\theta_n - (\delta_i - \eta_c + \lambda_{ic} + \tau_{ij}))}{\sum_{h=0}^{m_i} \exp \sum_{j=0}^{h} (\theta_n - (\delta_i - \eta_c + \lambda_{ic} + \tau_{ij}))} \quad x_i = 0, 1, 2, \ldots, m_i \qquad (3)$$

For the purpose of measuring parameter equivalence across a group of national subsamples c, an additional parameter for national effects on the item parameter λ_{ic} (the item-by-country interaction) is added to the model. However, to obtain proper estimates, it is also necessary to include the overall national effect (η_c) in the model.[3] Both item-by-country interaction estimates (λ_{ic}) and overall country effects (η_c) are constrained to having a sum of 0.

An even less constrained model for polytomous items would also have a country interaction parameter and, instead of the term τ_{ij}, an interaction effect between country and step parameters τ_{ijc} (see, e.g., in Walker, 2007). Such a model allows the estimation of separate step parameters for each country. As reviewing and interpreting the results of such an analysis becomes rather cumbersome with larger numbers of national samples, only the item-by-country interaction effect was estimated in the ICCS field trial analyses.

In this article, we illustrate how Rasch modelling can be used for testing assumptions about measurement equivalence in international studies. Based on the Rasch model (Rasch, 1960) for dichotomous items and the partial credit model for polytomous items (Wright & Masters, 1982), item-by-country interactions were estimated that provided useful information about the level of measurement invariance in parameter estimates.

Using example data from the IEA International Civic and Citizenship Education (ICCS), we will outline the following applications of Rasch modelling to a review of measurement invariance in the context of a cross-national study:

- using graphical information based on Rasch modelling and classical item statistics to review the performance of test items in country samples in cooperation with their respective national centres;
- comparing difficulty parameters for particular test items across national samples to detect cross-national item DIF as a criterion for the selection of material for the main data collection;
- assessing overall measurement equivalence for questionnaire Likert-type items designed to measure attitudinal traits with a facet Rasch model;
- using country-level indicators of questionnaire item DIF to review the occurrence of cross-national differences in measurement properties.

It should be acknowledged that within the framework of IRT there are alternative approaches to the analysis and treatment of measurement equivalence in cross-national studies. Rost's *mixed Rasch model* uses estimates of different sets of model parameters for different subpopulations to measure the same latent trait provided that the distribution of item difficulties is not substantially different (Rost, 1991; Rost, Carstensen, & von Davier, 1997). Another example is Fox's multilevel IRT model (Fox, 2005), which allows the modelling of different item parameters at lower levels (see, e.g., in Park & Bolt, 2008).

Within the context of the ICCS field trial analyses, it was not feasible to use more complex scaling models. For the selection of main survey item material, it was important to use relatively simple criteria and communicate effectively with national centre staff who may have limited or no knowledge of scaling methodologies. It was also necessary to ensure a considerable degree of transparency of selection procedures which would have been considerably reduced when using more complex modelling strategies.

Furthermore, using uni-dimensional Rasch models for assessing scaling properties of test items cross-nationally has become a standard in international studies of educational outcomes and is extensively used to communicate item results for national countries in comparison with those at the international level (Martin, Kennedy, & Trong, 2007; Olson et al., 2008; OECD, 2009).

Example data

The IEA International Civic and Citizenship Education Study (ICCS) was the third international IEA study designed to measure context and outcomes of civic and citizenship education, and it was explicitly linked through common questions to the IEA Civic Education Study (CIVED), which was undertaken in 1999 and 2000 (Amadeo, Torney-Purta, Lehmann, Husfeldt, & Nikolova, 2002; Schulz & Sibberns, 2004; Torney-Purta, Lehmann, Oswald, & Schulz, 2001).

The ICCS assessment framework (Schulz, Fraillon, Ainley, Losito, & Kerr, 2008) outlined the aspects that were addressed in the cognitive test and student perceptions questionnaire and provided a mapping of factors that might influence outcome variables and explain their variation. The main data collection took place between October and December 2008 in education systems with a Southern Hemisphere school calendar year and between February and May 2009 in those with a Northern Hemisphere school calendar year.

The study surveyed students in their eighth year of schooling from 38 countries and reported on students' civic knowledge, engagement, and perceptions, as well as on the contexts for civic and citizenship education (see Schulz, Ainley, Fraillon, Kerr, & Losito, 2010a, 2010b). The survey was designed 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 included an on-line national contexts survey completed by national centres, a student test, a student questionnaire, a teacher questionnaire, and a school questionnaire (see technical details in Schulz, Ainley, & Fraillon, 2011).

The following verification procedures were implemented prior to the international field trial and the main survey to ensure the 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 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, national centres submitted all translated instruments to be verified by professional language experts. 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 instruments and submitted them for final layout verification by the ISC. The results of this final check were sent back to the countries for final adjudication.

The data included in the analyses were collected during the ICCS field trial which was carried out in late 2007 and early 2008 and comprised data collections in 718 schools in 31 countries and comprised questionnaire data from 19,369 students, 9,383 teachers, and 681 school principals.[4]

The following international instruments[5] were used in the ICCS field trial:

- The international student test with 98 items was administered in a completely rotated design with six randomly allocated booklets, each consisting of three 20-min clusters.
- The international student questionnaire (with a total of 71 background and 201 perceptions items) was administered in three randomly allocated questionnaire forms.
- The international teacher questionnaire contained around 32 questions that took about 30 min to answer.
- The international school/principal questionnaire contained 22 questions which took 20 to 30 min to answer.

For the analysis of ICCS field trial data presented in this article, we used the ACER ConQuest software (Wu, Adams, Wilson, & Haldane, 2007) for the scaling of items with the Rasch model. Further analyses of field trial items included reviews of item frequencies, proportions of missing data, factor analyses, and correlations with related variables.

Calibrations of the ICCS field trial test items based on the pooled international sample suggested that the ICCS test items covered the range of student abilities found in the field trial sample. The test had a reliability coefficient (ACER ConQuest IRT estimate) of 0.86 for the pooled international sample, which shows that the ICCS field trial test was highly reliable. The results also showed a generally good item fit for the majority of ICCS test items. About 10% of the item material showed some form of item misfit, which was assessed based on several criteria including item fit indices, item-total correlations, and evidence of differential item functioning by gender or country.

The ICCS field trial questionnaire data were analysed using a wide range of statistical methods including both confirmatory factor analyses (CFA) and item response modelling (see Schulz, 2009). Most of the scales derived from the ICCS field trial data had satisfactory reliabilities above 0.70 (Cronbach's alpha); however, several of the trialled item sets were not retained for the ICCS main survey due to unsatisfactory scaling properties.

Results

Review of measurement invariance for test items

To ensure cross-national comparison of items, national calibration results were compared to those for the pooled international sample. To this end, we undertook separate calibrations of national subsamples as well as for the combined field trial dataset. The results were stored in a database that contained a wide range of international and national item statistics including category percentages, point biserials[6] for each item category, fit indices (weighted mean square statistic), the point-biserial correlation for the scored item (discrimination index), the parameter estimates item difficulty, and the Thurstonian thresholds.[7]

Based on this information, national study centres were provided with country reports designed to inform how the ICCS test items functioned in their respective national version and to encourage national centre staff to re-check flagged items for possible adaptation, translation, or layout/printing errors that were not detected during the verification procedures.

Figure 1 shows an example of a graphical display of item information for a particular ICCS test item based on a national calibration. At the top of the graph, the frequencies of item responses for this item in the national subsample were displayed. The upper bar chart depicts the average abilities (IRT weighted likelihood estimates in logits) within each item category, whereas the lower bar chart indicates the magnitude of point-biserial correlations for each item category.

The displayed graphs facilitated the process for identifying the following possible anomalies:

- A non-key category (incorrect or missing response) had a positive point-biserial; or a non-key category had a point-biserial higher than the key category.
- The correct (key) category had a negative point-biserial.
- In the case of scored partial credit items, checks could be made on whether the average ability (and the point-biserial) increased with the score points.

The example in Figure 1 is a multiple-choice item, where Category 2 represents the key, Categories 1, 3, and 4 the incorrect options, and Category 9 the missing responses. In this particular example, the key Category 2 had the highest average

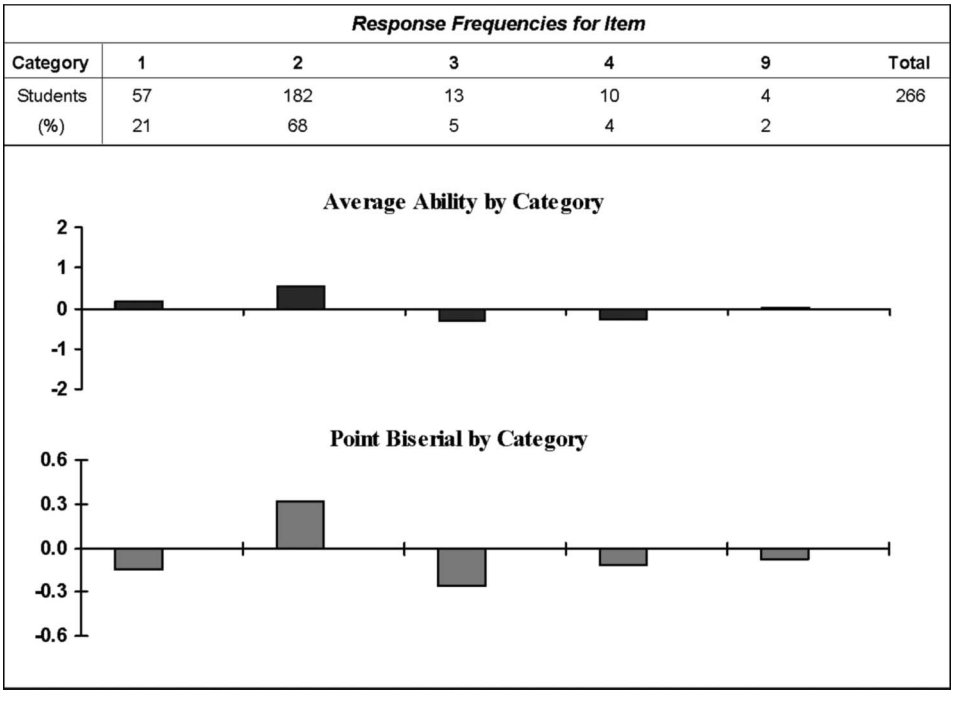

Figure 1. Example of item statistics* in graphical form.
*ACER ConQuest estimates.

ability estimates and was positively correlated with the overall score. All non-key responses had lower average abilities and also negative point-biserial item-total correlations.

Whereas the national item plots provided information about the particular fit of a particular item in each national subsample, further graphs presented comparisons of results between the national and pooled international sample. Substantial differences between the national value and the international value or the national value mean indicated that an item was measuring achievement differently in that particular country compared to other countries. Such a finding could be due to some technical problem (incorrect translation or printing error) or suggest a lack of appropriateness within the particular national context. However, if the item had poor scaling properties in all or nearly all countries, it suggested a general problem with the international source item.

Example results of comparisons between national and international calibrations are graphically presented in Figures 2 and 3. For each item, all values obtained for the national sample were compared with the values computed for all other countries and the respective value at the international level. Black crosses represented the value of the coefficients computed from the international database and the grey boxes the average for the pooled international sample plus and minus one standard deviation for the overall distribution of national values. The crosses in grey colour indicated the respective values computed for the subsample of the country for which the report was returned.

Figure 2 shows an example of a graphical comparison between the national and international fit indices (weighted mean square statistic) and between the national and international item-total correlations. Larger differences suggested that items performed differently in the respective national subsample and would have warranted a more detailed review of the national statistics for this particular item (as those displayed in Figure 1).

Figure 3 illustrates the graphical comparison between national item difficulties and item-category thresholds and the international ones. Larger differences indicated

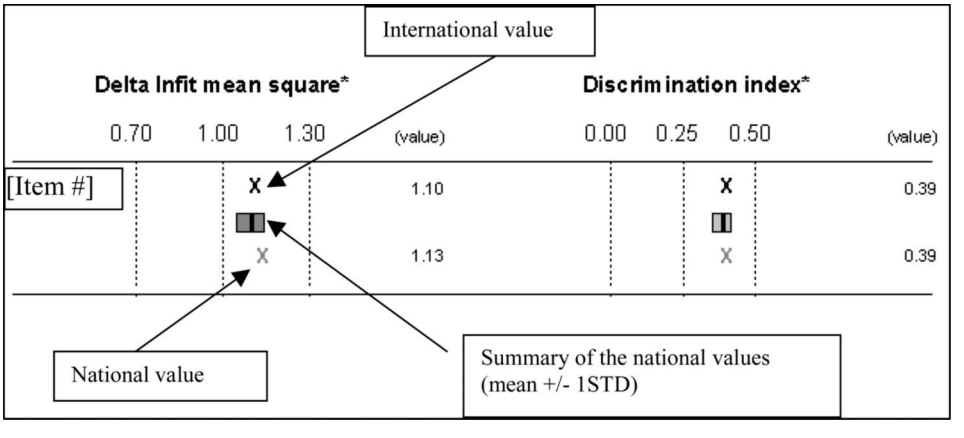

Figure 2. Example comparison of national and international item fit indices and item-total correlations.
* *ACER ConQuest estimates.

potential violations of the assumption of measurement invariance for the particular item with respect to the national subsample.

A list of items for which the field trial data had shown problems was sent to each national centre. These items were flagged if any of the following problems had been observed for the national version of an ICCS test item:

- The national item difficulty was lower than on average, and this difference was statistically significant at $p < 0.05$.
- The national item difficulty was higher than on average, and this difference was statistically significant at $p < 0.05$.
- One of the non-key categories had a point-biserial correlation higher than 0.05 (only reported if the category was chosen by at least 10 students).
- The item-total correlation was lower than 0.2; and/or
- the average abilities within categories for partial credit items were not ordered.

Figure 4 shows an example of a national item review list. The list summarised the observations for items that national centres were asked to revise. If an item turned out to be easier or harder than expected, national centres were asked to review (if possible in cooperation with national experts) the translation and also consider alternative explanations for these findings (e.g., curriculum, specific national context, recent events related to item content).

Whereas the comparison of national item statistics with those for the pooled international sample was designed to highlight particular problems for the national versions of the international source version, it was also important to assess the overall extent of item-by-country interactions for particular items.

To review the overall extent of item-by-country interactions, national calibrations of cognitive test item difficulties were plotted against those for the pooled international sample. To make the calibrated parameters comparable, all item parameters were standardised to having an average of zero for each calibration (both

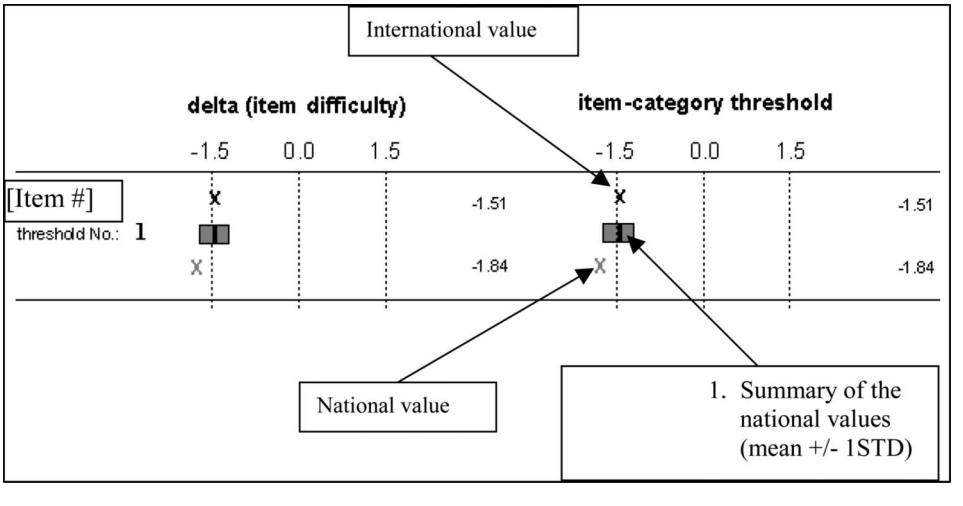

Figure 3. Example comparison of national and international item difficulties and thresholds.
* *ACER ConQuest estimates.

nationally and internationally). The results of these analyses are displayed in graphical form (see Figures 5 and 6).

For each item used in the cognitive test, there was one bar chart where the bars indicate the estimated item parameters for each participating country with their respective confidence intervals.[8] The dotted horizontal line indicates the estimated international item parameter. Bars that did not cross the dotted line indicated significant item-by-country interaction for a country and its respective item parameter. Example A in Figure 5 shows an item in which there is relatively little item-by-country interaction.

	Item by Country Interactions			Discrimination		
	No of Valid Responses	Easier than Expected	Harder than Expected	Non-key PB is Positive	low discrimi-nation	Ability not Ordered
[Item #1]	2226	☑	☐	☐	☐	☐
[Item #2]	2172	☑	☐	☐	☐	☐
[Item #3]	2193	☐	☑	☐	☐	☐
[Item #4]	2202	☐	☑	☐	☐	☐
[Item #5]	2207	☐	☑	☐	☐	☐

Figure 4. Example of national item review list.
* *Flagging based on ACER ConQuest estimates.

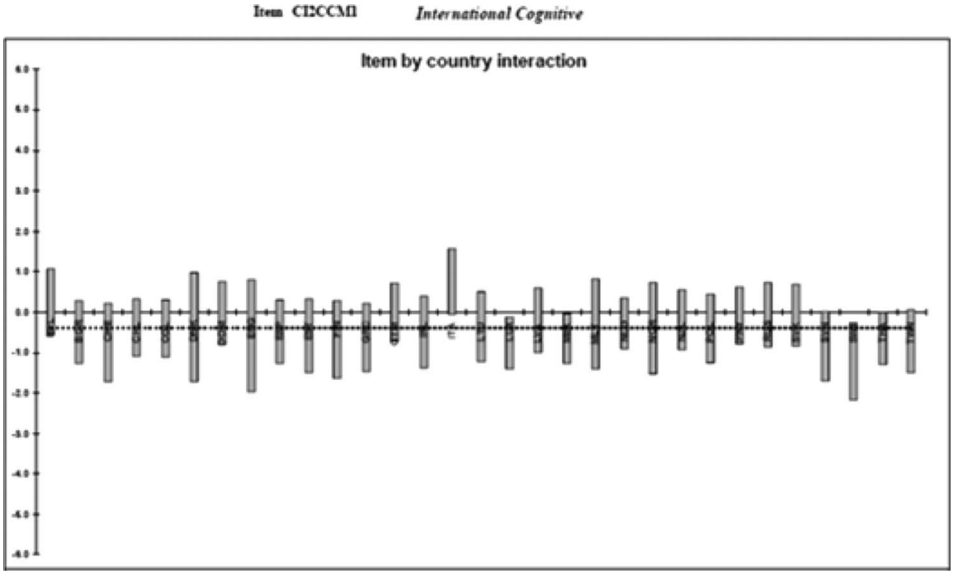

Figure 5. Graphical display of item-by-country interaction (example A).
* *ACER ConQuest estimates.

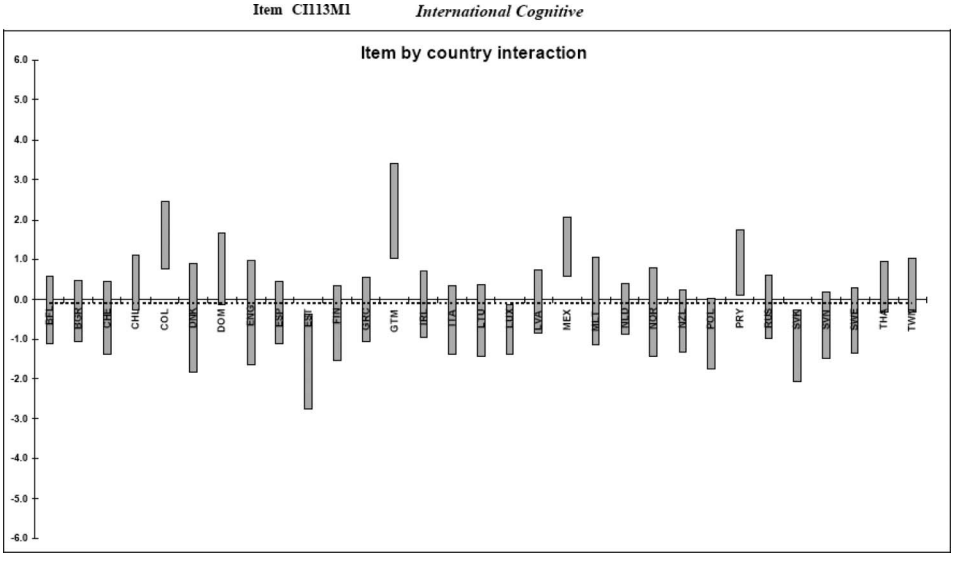

Figure 6. Graphical display of item-by country interaction (example B).
* *ACER ConQuest estimates.

Example B in Figure 6 shows the results for an item for which there was notable evidence of systematic interaction by geographical region. In this example, the results indicated that, after controlling for ability, the item was considerably harder in four of the six Latin American countries. This finding suggested that some part of the linguistic, sociocultural, or educational context common to these countries might have affected the relative difficulty of this particular test question.

Generally, only a minority of ICCS field trial test items showed item-by-country interactions. Field trial information about occurrence of cross-national DIF was used as one important criterion in the final selection of items for the main survey test instrument. However, it should be noted that occurrence of item-by-country interactions in a smaller number of countries did not automatically lead to the exclusion of field trial items from the material for the final ICCS civic knowledge test.

Review of measurement invariance for questionnaire items

To illustrate the review of measurement invariance for questionnaire items, we used a set of items designed to measure students' attitudes towards their country. Table 1 shows the items that were included in the ICCS field trail. Factor analyses had indicated that item IS2H02D (preferring to live in another country) did not fit the scale, and it was discarded from further scaling analyses.

Rasch modelling with the partial credit model (PCM) provided a particularly useful tool for reviewing measurement equivalence of questionnaire items across national subsamples. To this end, models with country interaction effects (see Equation (3)) were estimated that provided indicators of the degree of parameter invariance across national subsamples. The estimates (λ_{ic}) are on the same (logit) scale as item and person parameters and give an approximation of how much a

nationally calibrated item location parameter would have deviated from the corresponding estimate based on the pooled international sample.

To describe the degree of parameter variation across countries for each item, it was deemed important to summarise this information. In the ICCS field trial analysis, the median of absolute values for item-by-country interaction effect was taken as an indicator of overall parameter invariance for each item. The median was chosen as it is less affected by outliers. In addition, the minimum and maximum effects were displayed to demonstrate the absolute range of deviations across countries.

Table 2 shows the results from the international calibration using the pooled field trial sample as well as summary indicators of item-by-country interactions for each of these items. The results suggested that in particular items IS2H02F and IS2H02H did not fit well with the overall scale. In terms of item fit, both items were less discriminating than others and also had somewhat lower item-total correlations. The summary of country DIF indicated that in particular the items IS2H02A and IS2H02G showed higher median values of item-by-country interaction as well as a wider range of these estimates.

Table 3 shows the estimated item-by-country interaction terms for the item location parameter. Positive values indicated that a nationally calibrated item parameter would have been estimated as relatively harder to agree with than in the international calibration. Negative values indicated that the corresponding national calibration would have rendered a lower location parameter than in the pooled international sample.

Item IS2H02A (''importance of flag'') had the highest median country item-by-country interaction. The country-level results revealed that students in many developed countries (Northern and Central Europe, New Zealand) found it relatively harder to agree with this item, whereas, after controlling for the latent trait, students from Latin American or Southern European countries showed a tendency to agree more readily with this item. In other words, the location parameter for this item would have been quite different when calibrating in Latin America compared to those from a calibration in Scandinavian countries. To a lesser extent,

Table 1. Items measuring attitudes towards own country.

Item	Item wording	Used for scaling?
IS2H02A	The <flag of country of test> is important to me	Yes
IS2H02B	I have great respect for <country of test>	Yes
IS2H02C	In <country of test> we should be proud of what we have achieved	Yes
IS2H02D	I would prefer to live permanently in another country	No
IS2H02E	I am proud to live in <country of test>	Yes
IS2H02F	People should support <country of test> even if its government is doing the wrong thing	Yes
IS2H02G	Generally speaking, <country of test> is a better country to live in than most other countries	Yes
IS2H02H	The world would be a better place if citizens from other countries were like <citizens of country of test>	Yes

Note: Expressions in < > were adapted by national centres. Response categories were ''strongly agree'', ''agree'', ''disagree'', and ''strongly disagree'', coded as 3, 2, 1, and 0 for scaling purposes.

there was a similar pattern for item IS2H02B (“great respect for country”), which was another item based on a statement endorsing symbolic patriotism.

We also found considerable item-by-country interaction for item IS2H02G (“own country better to live in than others”). This item tended to be relatively easier to endorse in developed countries (particularly in Northern Europe), whereas students in Latin American and Eastern European countries found it relatively harder to agree with. This finding suggests that, as a result of separate calibrations of country subgroups, a considerably lower item location parameter would have been estimated in wealthier countries than in poorer countries. Item IS2H02H (“world better place if all were like country’s citizens”), another item for which responses were likely to be affected by the general economic wealth of the students’ country, showed a similar pattern, but there was generally less item-by-country interaction for this item.

In summary, the results showed that there was a noticeable lack of measurement equivalence for the trialled items measuring students’ attitude towards their own country. It is interesting to note that those items related to symbolic patriotism were relatively easier to agree with in Latin American and Mediterranean countries but relatively harder to be endorsed in Northern European countries. Not unexpectedly, items that may be more affected by objective living conditions were found to be relatively harder to agree with in poorer countries.

Information like the one presented in this example was used as one important criterion for the item selection for the ICCS main survey, and preference was given to items that had less item-by-country interaction. However, as in the example described in this section, differential behaviour of items can also be quite informative and point out interesting patterns across countries.

Table 2. IRT summary table for items measuring attitudes towards own country (calibration results and summary of country DIF).

		International Calibration results			Item-by-country interaction		
Item	Content	*Parameter*	*Fit*	Item-total correlations	*Median of absolute values*	*Minimum*	*Maximum*
IS2H02A	Flag important	–0.28	0.98	0.70	0.59	–1.00	1.18
IS2H02B	Great respect for country	–0.65	0.83	0.76	0.26	–1.01	0.68
IS2H02C	Proud of achievement	–0.59	0.86	0.75	0.11	–0.91	0.75
IS2H02E	Proud to live in country	–0.48	0.84	0.77	0.11	–0.42	0.44
IS2H02F	Support country always	0.59	1.25	0.63	0.27	–0.74	0.96
IS2H02G	Better country than others	0.35	1.10	0.67	0.48	–0.90	1.15
IS2H02H	World better place	1.06	1.22	0.64	0.25	–0.58	0.93

Note: ACER ConQuest estimates.

Table 3. IRT item-by-country interactions for items measuring attitudes towards own country.

	Flag important	Great respect for country	Proud of country"s achievement	Proud to live in country	Support country always	Better country than others	World better place
Country	IS2H02A	IS2H02B	IS2H02C	IS2H02E	IS2H02F	IS2H02G	IS2H02H
Belgium (Flemish)	**1.18**	0.09	0.13	0.03	***–0.64***	***–0.38***	***–0.41***
Bulgaria	***–0.58***	***–0.48***	–0.07	0.09	–0.16	**0.73**	**0.45**
Chile	***–0.39***	***–0.32***	–0.02	–0.01	0.27	0.22	0.25
Chinese Taipei	–0.19	**0.53**	0.18	**0.35**	–0.06	***–0.34***	***–0.48***
Colombia	***–0.57***	***–0.44***	0.05	***–0.33***	**0.69**	**0.50**	0.10
Denmark	**0.77**	**0.48**	–0.12	0.16	***–0.44***	***–0.62***	–0.24
Dominican Republic	***–0.89***	0.13	0.01	0.05	**0.96**	0.15	***–0.40***
England	**0.78**	0.13	–0.16	***–0.34***	0.18	***–0.32***	–0.28
Estonia	0.02	–0.03	–0.24	0.16	–0.27	0.02	**0.34**
Finland	0.10	**0.42**	0.11	0.05	–0.21	***–0.72***	0.25
Greece	***–0.93***	***–0.47***	–0.03	0.10	–0.06	**0.49**	**0.90**
Guatemala	***–0.36***	–0.07	–0.15	–0.06	–0.19	**0.49**	**0.33**
Ireland	0.10	***–0.30***	***–0.43***	***–0.38***	**0.40**	0.23	**0.39**
Italy	***–0.71***	***–0.30***	**0.75**	–0.18	0.09	0.13	0.21
Latvia	***–0.36***	0.25	0.23	**0.37**	***–0.36***	0.12	–0.25
Lithuania	***–0.76***	***–1.01***	***–0.91***	–0.11	**0.70**	**1.15**	**0.93**
Luxembourg	**0.69**	**0.68**	0.18	0.12	***–0.50***	***–0.76***	***–0.41***
Malta	–0.17	0.11	–0.11	0.14	–0.03	0.04	0.04
Mexico	***–0.70***	–0.26	–0.03	–0.09	**0.49**	**0.48**	0.12
Netherlands	**1.08**	0.27	0.14	0.14	***–0.35***	***–0.70***	***–0.58***
New Zealand	**0.83**	0.11	–0.03	–0.15	–0.06	***–0.32***	***–0.39***
Norway	**0.74**	0.12	–0.01	–0.09	0.03	***–0.90***	0.10
Paraguay	***–1.00***	***–0.62***	0.06	0.14	**0.94**	**0.48**	0.00
Poland	***–0.59***	–0.25	0.25	**0.44**	***–0.74***	**0.68**	0.20
Russian Federation	0.21	0.08	0.08	–0.10	–0.14	0.09	–0.22
Slovak Republic	0.09	0.02	–0.13	–0.03	***–0.70***	**0.78**	–0.05
Slovenia	0.15	**0.41**	0.09	0.02	–0.11	–0.22	***–0.34***
Spain	0.18	0.15	0.17	***–0.42***	0.17	–0.15	–0.09
Sweden	**0.82**	**0.59**	–0.05	–0.09	***–0.46***	***–0.60***	–0.21
Switzerland	**0.64**	0.19	0.10	–0.02	–0.07	***–0.74***	–0.10
Thailand	–0.20	–0.21	–0.05	0.04	**0.62**	–0.04	–0.17

Note: Item-by-country interaction > 0.3 logits are highlighted in **bold** and those < –0.3 in ***bold italics***.

Discussion and conclusion

The Rasch IRT model provided an invaluable tool for assessing the extent of measurement equivalence in the ICCS field trial. For cognitive test items, scaling properties were reviewed separately for each national sample and compared with the overall calibration results. Through graphical communication, it was possible to detect and display possible item problems to national centres that were subsequently re-checked. The general variation of estimated parameters across participating countries provided an important indicator for the final selection of test item material.

Another important feature of the analysis procedures for the ICCS field trial was the implementation of an IRT analysis of measurement equivalence with questionnaire material.

Some aspects regarding cross-national comparability raised in the literature could not be addressed within the scope of the ICCS field trial analyses. For example, some researchers have raised concerns with regard to the appropriateness of using Likert-type items for measuring constructs in cross-cultural studies because of differences in response patterns across countries (see, e.g.,, Heine, Lehman, Peng, & Greenholtz, 2002). The review of parameter invariance could also be further extended to the functioning of step parameters to investigate differences in cross-national response patterns for Likert-type questionnaire items (see, e.g., Walker, 2007).

It should be noted that the focus of decisions about the degree of item-by country interaction in our analysis of ICCS field trial data had been on the relative scaled item difficulties between individual countries and the international scale. Information about item-by-country interaction was part of the evidence regarding the item functioning but did not automatically determine the item selection process. Ultimately, when selecting items for the final instruments, a balance had to be struck between a level of item variability that would not compromise the integrity of the instrument measures and the necessity to have sufficient items to cover the breadth of the conceptual framework underlying the assessment. Fortunately, overall the ICCS field trial data and analysis revealed only a small number of test items that showed higher levels of item-by-country interaction.

The analyses undertaken with ICCS field trial data using Rasch IRT modelling showed a noticeable but limited amount of item-by-country interaction. It should be noted that stringent tests of measurement equivalence would routinely lead to the rejection of items due to the large sample sizes typically obtained in international survey studies. Therefore, ICCS field trial data on parameter invariance were rather interpreted as relative measures, and ''rules of thumb'' had to be developed to flag higher levels of parameter variation.

The examples shown in this article illustrate how item parameters may vary depending on the context of participating countries and how this information may be used prior to the final item selection at the field trial stage of cross-national studies. Item parameter variation is not necessarily the result of a specific technical problem with an item. It can also be the result of genuine measured differences between groups (countries in the case of ICCS) and may indeed provide valuable information that should not automatically be dismissed as ''undesirable measurement bias''.

Careful analyses of the item properties and the known contexts and attributes of subgroups need to be undertaken to decide how to manage items showing parameter variation. Reviewing measurement equivalence is important in comparative research, but it should not lead to a simple ''one-size-fits-all'' approach that may exclude many interesting aspects from educational research and that would reduce comparisons on those issues that are relatively uniform across countries. Instead, it is our view that ''on-balance'' judgments are warranted that preserve interesting differences without jeopardising the validity of cross-national measures.

When analysing parameter invariance in international studies, it is also possible to model the occurrence of cross-national DIF as a case of multidimensionality in the Rasch model. For example, variation in parameter estimates may be explained by factors like language (see Grisay, Gonzalez, & Monseur, 2009) or curriculum

coverage (Hencke, Rutkowski, Neuschmidt, & Gonzalez, 2009). One important question in this context which still needs further exploration is at what point parameter invariance starts making a real difference and leads to bias when it comes to measuring constructs in cross-national studies.

Notes

1. It should be noted that the Rasch model does not make any adjustments for differences in item discrimination or guessing of responses to multiple-choice items as is the case when using the IRT two- or three-parameter models where this is explicitly modelled through the inclusion of additional item parameters (see Hambleton et al., 1991).
2. Item characteristic curves are graphical displays that plot the expected against the observed item responses. In addition, these plots can be used to assess the responses to distractor categories (incorrect options) in a multiple-choice item.
3. The minus sign ensures that positive values of the country group effect parameters indicate relatively higher levels of item endorsement in a country.
4. One national centre submitted its data at a later stage, and its data could not be included in the analyses.
5. 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.
6. *Point biserials* are correlation coefficients between dichotomous indicator variables for each item category and the overall score.
7. For polytomous items scaled with the Rasch partial credit model, Thurstonian thresholds indicate for each item category those points on a scale at which respondents have a 0.5 probability of scoring this particular category or higher. In the case of items with only two categories scaled with Rasch model for dichotomous items, they are equivalent to the item difficulty or location parameter.
8. The confidence intervals for item parameters were adjusted for design effects (due to the cluster sample design) and multiple comparisons (for 31 countries).

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, The Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

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, CT: 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, The Netherlands: IEA.

Fox, J.P. (2005). Multilevel IRT model assessment. In L.A. van der Ark, M.A. Croon, & K. Sijtsma (Eds.), *New developments in categorical data analysis for the social and behavioral sciences* (pp. 227–252). Mahwah, NJ: Lawrence Erlbaum.

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, France: OECD Publications.

Grisay, A., Gonzalez, E., & Monseur, C. (2009). Equivalence of item difficulties across national versions of the PIRLS and PISA reading assessments. In *IERI monograph series: Issues and methodologies in large-scale assessments*. (Vol. 2, pp. 63–83). Retrieved from http://www.ierinstitute.org/IERI_Monograph_Volume_02.pdf

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

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

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*, 903–918.

Hencke, J., Rutkowski, L., Neuschmidt, O., & Gonzalez, E. (2009). Curriculum coverage and scale correlation on TIMSS 2003. In *IERI monograph series: Issues and methodologies in large-scale assessments* (Vol. 2, pp. 85–112). Retrieved from http://www.ierinstitute.org/IERI_Monograph_Volume_02.pdf

Jak, S., Oort, F.J., & Dolan, C.V. (2010). Measurement bias and multidimensionality: An illustration of bias detection in multidimensional measurement models. *ASTA Advances in Statistical Analysis*, *94*, 129–137.

Martin, M.O., Kennedy, A.M., & Trong, K.L. (2007). Item analysis and review. In M.O. Martin, I.V.S. Mullis, & A.M. Kennedy (Eds.), *PIRLS 2006 technical report* (pp. 131–148). Amsterdam, The Netherlands: IEA.

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, NY: Springer.

Olson, J.F., Martin, M.O., & Mullis, I.V.S. (2008). *TIMSS 2007 technical report*. Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

Olson, J.F., Martin, M.O., Mullis, I.V.S., Foy, P., Erberber, E., & Preuschoff, C. (2008). Reviewing the TIMSS 2007 item statistics. In J.F. Olson, M.O. Martin, & I.V.S. Mullis (Eds.), *TIMSS 2007 technical report* (pp. 193–224). Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

Organisation for Economic Co-operation and Development. (2009). *PISA 2006 technical report*. Paris, France: Author.

Park, C., & Bolt, D.M. (2008). Application of multilevel IRT to investigate cross-national skill profiles on TIMSS 2003. In *IERI monograph series: Issues and methodologies in large-scale assessments* (Vol. 1, pp. 71–96). Retrieved from http://www.ierinstitute.org/IERI_Monograph_Volume_01.pdf

Perrone, M. (2006). Differential item functioning and item bias: Critical consideration in test fairness. *Teachers College, Columbia University Working Papers in TESOL and Applied Linguistics*, *6*, 1–3.

Raju, N.S., Laffitte, L.J., & Byrne, B.M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, *87*, 517–529.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: The Danish Institute of Education Research. (Expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago, IL: The University of Chicago Press).

Rost, J. (1991). A logistic mixture distribution model for polytomous item responses. *British Journal of Mathematical and Statistical Psychology*, *44*, 75–92.

Rost, J., Carstensen, C., & von Davier, M. (1997). Applying the mixed Rasch model to personality questionnaires. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 324–332). Münster, Germany: Waxmann.

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

Schulz, W. (2009). Questionnaire construct validation in the International Civic and Citizenship Education Study. In *IERI monograph series: Issues and methodologies in large-scale assessments* (Vol. 2, pp. 113–135). Retrieved from http://www.ierinstitute.org/IERI_Monograph_Volume_02.pdf

Schulz, W., Ainley, J., & Fraillon, J. (Eds.). (2011). *ICCS 2009 technical report*. Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

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

Schulz, W., Ainley, J., Fraillon, J., Kerr, D., & Losito, B. (2010b). *Initial findings from the IEA International Civic and Citizenship Study*. Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

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

Schulz, W. & Sibberns, H. (Eds.). (2004). *IEA Civic Education Study. Technical report*. Amsterdam, The Netherlands: 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, The Netherlands: 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*, 267–278.

Wright, B.D., & Masters, G.N. (1982). *Rating scale analysis: Rasch measurement*. Chicago, IL: MESA Press.

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, Victoria, Australia: ACER Press.