The Analysis of Measurement Equivalence in International Studies using the Rasch Model

Wolfram Schulz
Australian Council for Educational Research
Melbourne/Australia
schulz@acer.edu.au

Julian Fraillon
Australian Council for Educational Research
Melbourne/Australia
fraillon@acer.edu.au

Paper presented to the symposium on "Rasch measurement: present, past and future"
at the European Conference on Educational Research (ECER)
in Vienna, 28-30 September 2009.

Introduction

Most international studies such as TIMSS, PIRLS, PISA, CIVED or ICCS make use of item response modelling 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 scaling methods) in cross-national research is measurement invariance because typically international parameters are used for obtaining scale scores that are used for between-country comparisons.

It is widely recognised that language differences may have a powerful effect on equivalence (or non-equivalence) of test and questionnaire items and this consequently challenges assumptions made about measurement invariance. Most international studies (see for example Grisay, 2002; Chrostowski & Malak, 2004) implement 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. Furthermore, non-equivalence can also be caused by the cultural differences among participating countries in international studies (van de Veijver and Tanzer, 1997; Byrne, 2003).

This paper outlines how Rasch modelling can be used for testing assumption about measurement equivalence in international studies. Based on the Rasch model (Rasch, 1960) for dichotomous items and the Partial Credit model for polytomous items (Masters & Wright, 1982), so-called item-by-country interactions can be estimated that provide useful information about the level of measurement invariance in parameter estimates. It should be noted that item-by-country interactions are one form of differential item functioning (DIF) where groups vary in their probability of answering questions even after controlling for their levels of ability (see Hambleton, Swaminathan and Rogers, 1991). Item-by-country interactions can be obtained from comparing separate calibrations for national sub-samples or estimated directly through the inclusion of additional model parameters (see examples in Walker, 2007 and Schulz, 2009).

Using field trial data from the IEA International Civic and Citizenship Education Study (ICCS) this paper outlines ways of obtaining estimates of measurement invariance and discusses their interpretations as well as limitations of these analyses. In particular, the paper illustrates the analysis of item-by-country interactions for test items and the estimation of country effects when scaling questionnaire (Likert-type) items.

The International Civic and Citizenship Education Study (ICCS)

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 surveys 13-to-14-year old students in 38

countries in the years 2008 and 2009 and will report on students' civic knowledge, engagement and perceptions as well as on the context for 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. In addition, an on-line survey carried out through national centres will inform on the context of civic and citizenship education at the national level.[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 et. al., 2008) outlines the aspects that 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 took place between October and December 2008 in the educational systems with Southern Hemisphere school calendar year and between February and May 2009 in those with a Northern Hemisphere school calendar year.

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

[1] Further information about ICCS can be found at its website http://iccs.acer.edu.au/.

submitted them for 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 following international instruments were used in the ICCS 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 analyses presented in this paper will focus on examples from the ICCS field trial student test and questionnaire. Field trial data from regional instruments and from teacher and school questionnaires underwent similar analysis procedures.

Rasch Model and Applications

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

For multiple-choice items and short-answer items with a category scored 1 for correct responses and 0 for incorrect responses, this is modelled as

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

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_n.

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 and Wright, 1997) as

[2] One national centre submitted its data at a later stage and its data could not be included in the analyses.

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

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.

It should be noted that both item and person (ability) parameters are measured on the same scale: In the case of dichotomous items (θ_n) with just two categories ("correct" and "incorrect") students who have the same ability as the estimated difficulty parameter of an item (δ_i) their probability of giving a correct response is 0.5.

The goodness of fit for individual items can be determined by calculating a Mean Square Statistic (Wright and Masters, 1982). Reviewing this residual-based item fit indicates the extent to which each item fits the item response model. However, there are no clear rules for acceptable item fit, and it is generally recommended that analysts and researchers interpret residual-based statistics with caution (see Rost & von Davier, 1994). Therefore, additional indicators like Item Characteristic Curves (ICC) were used to assess item fit which consist of plots of the average percent of observed responses for groups with similar values on the latent variable against their expected values for an item and give additional graphical information on item fit.

When applying a parametric measurement model parameter invariance is assumed. The measurement model should neither vary across sub-groups within countries nor across countries. IRT scaling enables researchers to test so-called Differential Item Functioning (DIF), which consists of different measurement characteristics depending on sub-groups within a sample. For example, Gender DIF occurs when a test items is relatively easier for girls than for boys or vice versa. Likewise, cross-country DIF or "item-by-country interaction" occurs when an item is relatively easier in one or more countries compared to its difficulty than in other countries.

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

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

For the purpose of measuring parameter equivalence across a group of national sub-samples *c*, an additional parameter for national effects on the item parameter λ_{ic} (the item-by-country interaction) is added to the (constrained) 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 have also a country interaction and replacing the term τ_{ij} with an interaction between country and step parameters τ_{ijc} (see an example 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 sample only the item-by-country interaction effect was estimated in the ICCS field trial analyses.

Test Item Analysis with the Rasch Model

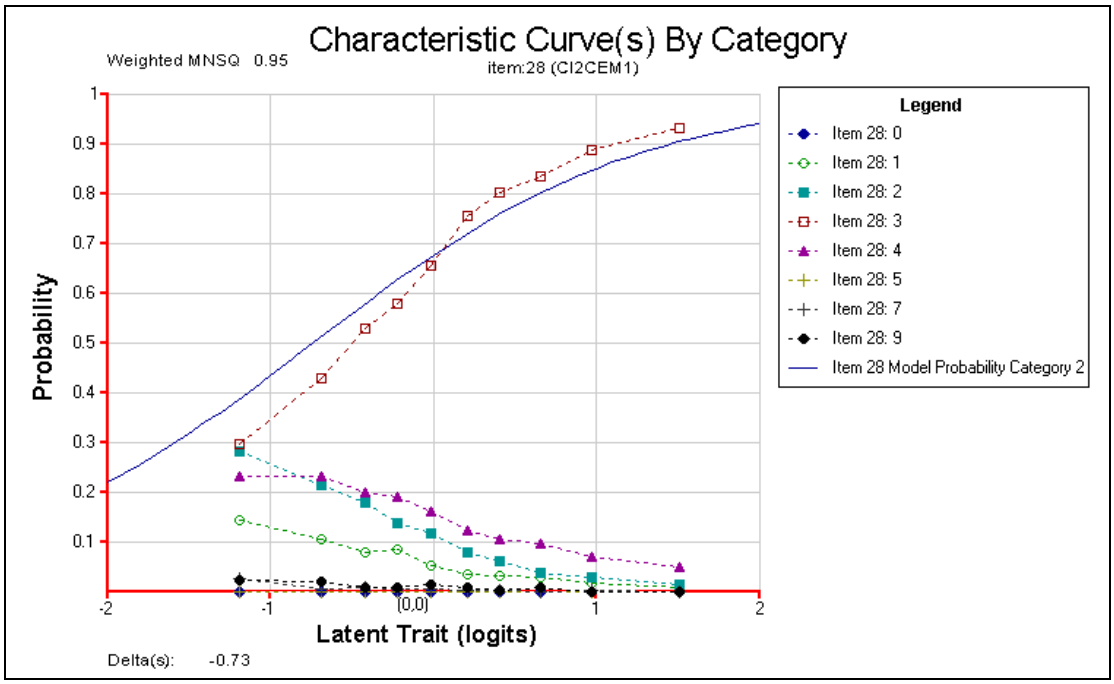
The field trial data were used to check the appropriateness of the test instrument with regard to different aspects including test length, item functioning, match between test difficulty and student abilities as well as the measurement equivalence of test items across educational systems.

Item response modelling using the ACER ConQuest software (Wu, Adams, Wilson, & Haldane, 2007) was applied to review the general match between test difficulty and student abilities and the scaling characteristics of individual items. Generally, the results indicated that the ICCS test items covered the range of student abilities in the field trial sample. The test had an overall reliability of 0.86 which shows that the ICCS field trial test is highly reliable.

The results for the pooled field trial sample showed that test items had an average discrimination of 0.37 and only ten out of 98 items were flagged for having a discrimination of less than 0.2. There were no items with a fit index below 0.88 and ten items with a weighted mean square statistic greater than 1.12. Items with poor discrimination or fit indices were substantively analysed for possible refinement and inclusion in the main survey instruments or removal from the item pool. Item functioning was also evaluated graphically by inspecting the item characteristic curve (ICC) for each item. ICC show graphically the actual student performance data on each item (represented by the response probability for the correct response across subgroups of students matched by ability) compared to the ideal function of response probability by student ability predicted by the Rasch model for each item. In addition to this information, ICC can be displayed to include the response probabilities of all response and score categories for multiple-choice and open ended response items. Figure 1 shows an example of an ICC for a multiple choice item that was deemed to be functioning well.

[3] The minus sign ensures that positive values of the country group effect parameters indicate relatively higher levels of item endorsement in a country.

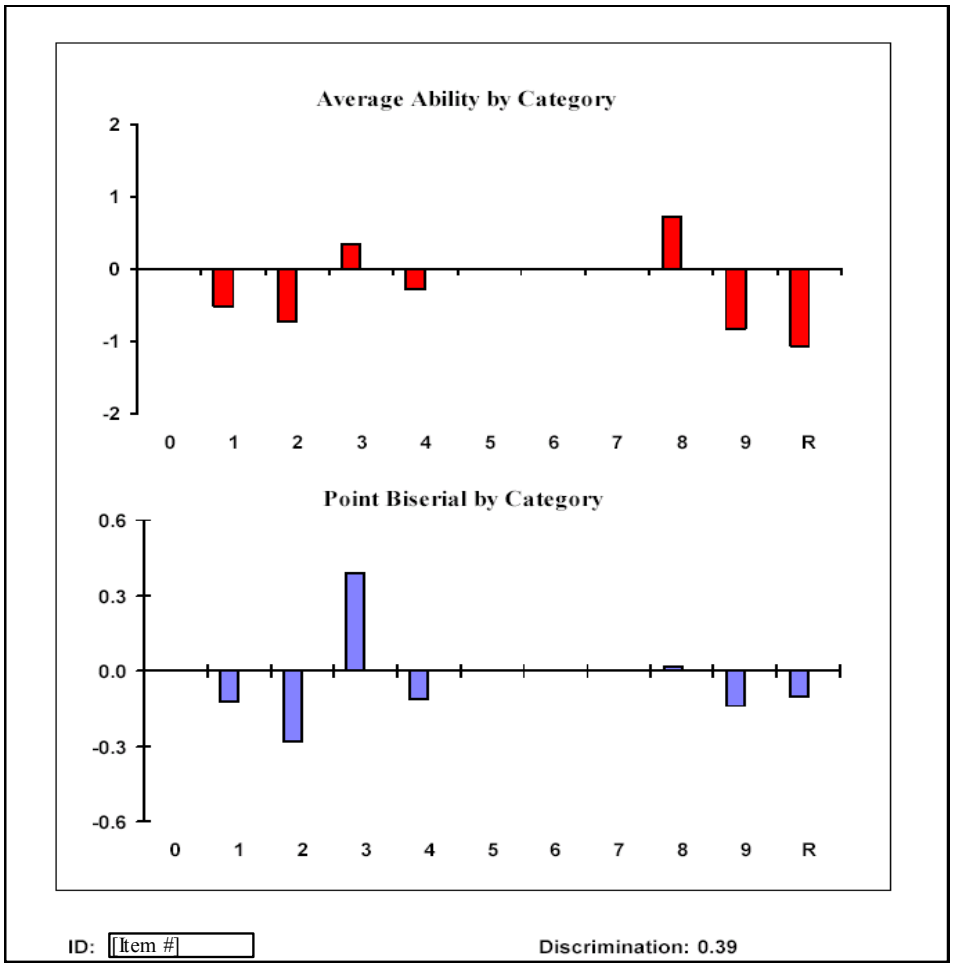
Figure 1: Example Item Characteristic Curve*



* The ICC was produced using ACER ConQuest.

To ensure cross-national comparison of items national calibration results were compared to those for the pooled international sample. For each educational system, graphical displays showed comparisons between category percentages, point biserials for each test item as well as estimated item fit, item discrimination and item difficulty for the national sub-sample compared to the pooled international sample. Additional tables flagged indicators of item misfit. These national reports were designed to inform national centres about how their particular test instrument worked and allow them to re-check flagged items for adaptation or translation errors that might not have been detected during the verification procedures.

Figure 2 Example of Item Statistics* in Graphical Form



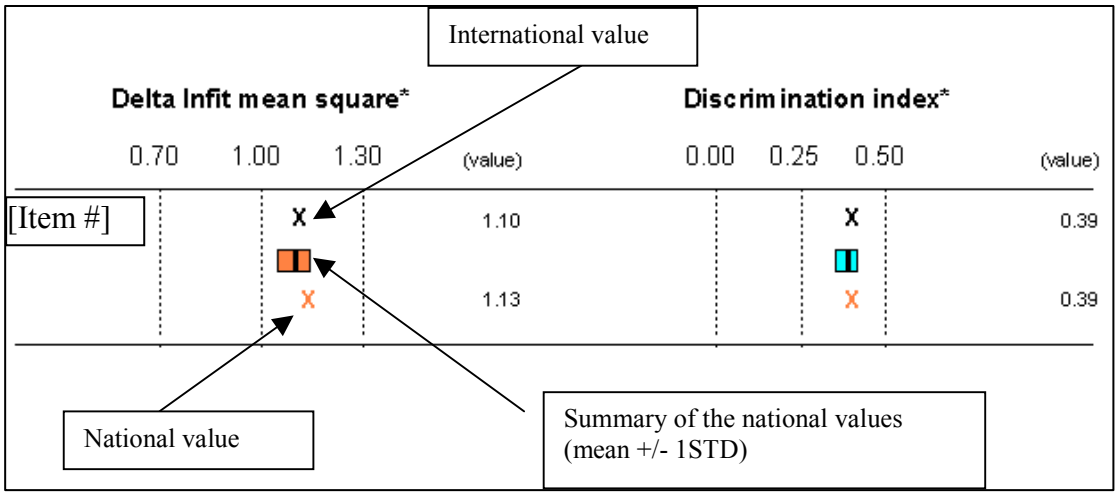
* ACER ConQuest estimates.

The average ability of IRT estimates and the point-biserial correlation by category were presented in a graphical form (see Figure 2). The average ability by category was calculated by domain, and centred for each item. The centring of the ability distribution allowed easy identification of "positive" and "negative" ability categories, so that checks could be made to ensure that, in the case of multiple-choice items, the key category had the highest average ability estimate; and for constructed items, that the mean abilities were ordered consistently with the score levels.

In addition, the displayed graphs facilitated the process for identifying the following possible anomalies:

- A non-key category had a positive point-biserial; or a non-key category had a point-biserial higher than the key category.
- The 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) increases with the score points.

Figure 3 Example of National and International Item Fit and Discrimination*



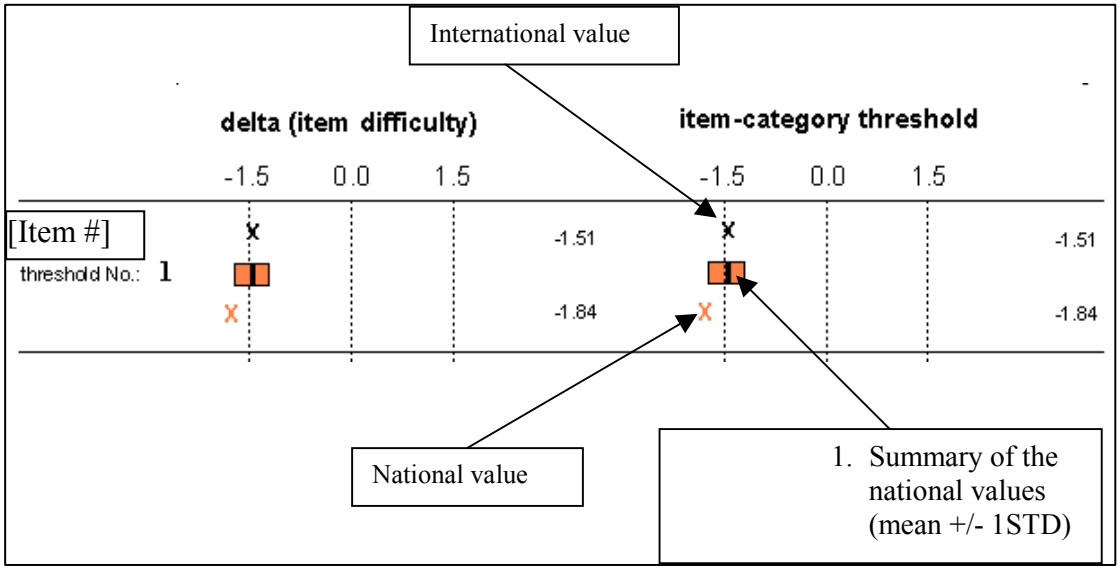
* ACER ConQuest estimates.

Calibrations for national sub-samples had provided for each country and for each item the Delta infit mean square, the point-biserial correlation, the Delta item parameter estimate (or difficulty estimate) and the thresholds. Example results are graphically presented in Figure 3. For each item the value computed for one country was compared with the values computed for all other countries and the value computed at the international level.

In Figure 2 the black crosses represent the value of the coefficients computed from the international database. The coloured boxes represent the mean plus and minus one standard deviation of these national values. The orange crosses represent the values computed only on a national data set, i.e. the data set of the country to which the report is returned.

Substantial differences between the national value and the international value or the national value mean could indicate that the item was behaving differently in that country in comparison with the other countries. Such a finding might reflect a mistranslation or another problem. On the other hand, if the item was misbehaving in all or nearly all countries, it might rather reflect a problem in the international source item.

Figure 4 Example of National and International Item Difficulties and Thresholds*



* ACER ConQuest estimates.

Substantial differences between the national value and the international value or the mean of national values for item fit or item discrimination indicated if an item behaved differently in a particular country. To assess such an "item-by-country interaction", 4 illustrated the differences between national item difficulties and item-category thresholds and the international ones.

A list of items which the field trial data had shown problems was sent to each national centre. These items were flagged if any of the following problems were observed:

- an item difficulty was significantly lower than on average;
- an item difficulty was significantly higher than on average;
- 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);
- an item discrimination[4] was lower than 0.2; and/or
- category abilities for partial credit items were not ordered.

[4] Item discrimination is defined here as the correlation of correct responses with the overall score.

Figure 5 Example of National Item Review List*

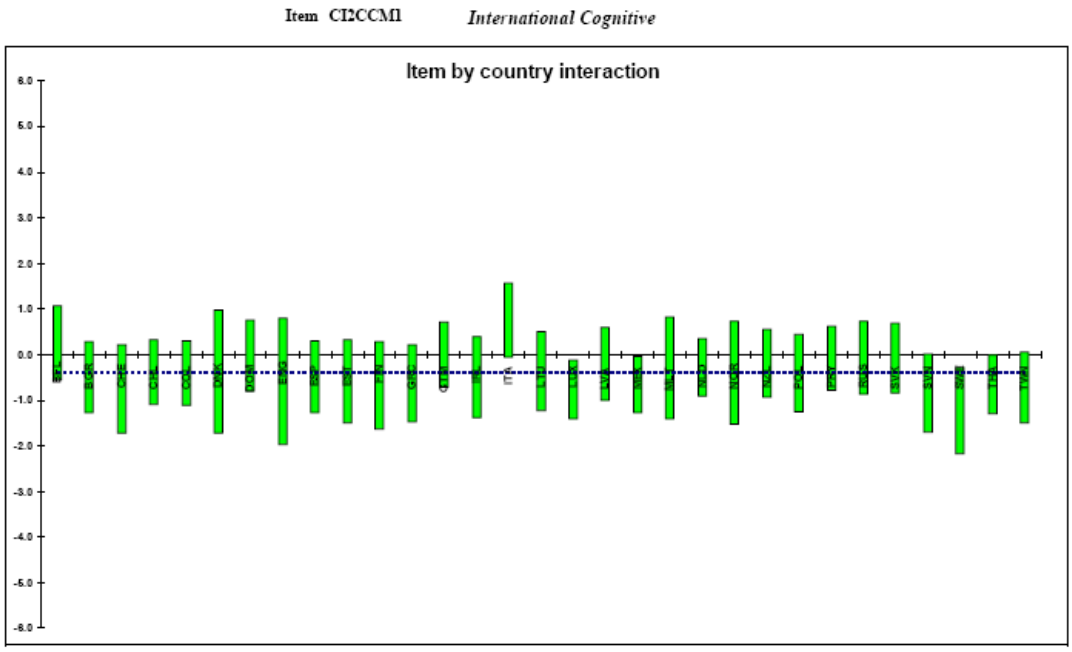
		Item by Country Interactions		Discrimination		
	No of Valid Responses	Easier than Expected	Harder than Expected	Non-key PB is Positive	low discrimi-nation	Ability not Ordered
[Item #1]	2226	☑	☐	☐	☐	☐
[Item #2]	2172	☑	☐	☐	☐	☐
[Item #3]	2193	☐	☑	☐	☐	☐
[Item #4]	2202	☐	☑	☐	☐	☐
[Item #5]	2207	☐	☑	☐	☐	☐

* Flagging based on ACER ConQuest estimates.

The "national item review list" in 5 summarises how items perform across countries that should be revised by the national centre. 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 (for example curriculum, specific national context, recent events related to item content).

To estimate item-by-country interactions, the cognitive test items were calibrated separately for each national sub-sample and for the pooled international sample. All item parameters were standardised to having an average of zero for each test (both nationally and internationally). The results of these analyses were displayed in graphical form (see Figure 6 and Figure 7).

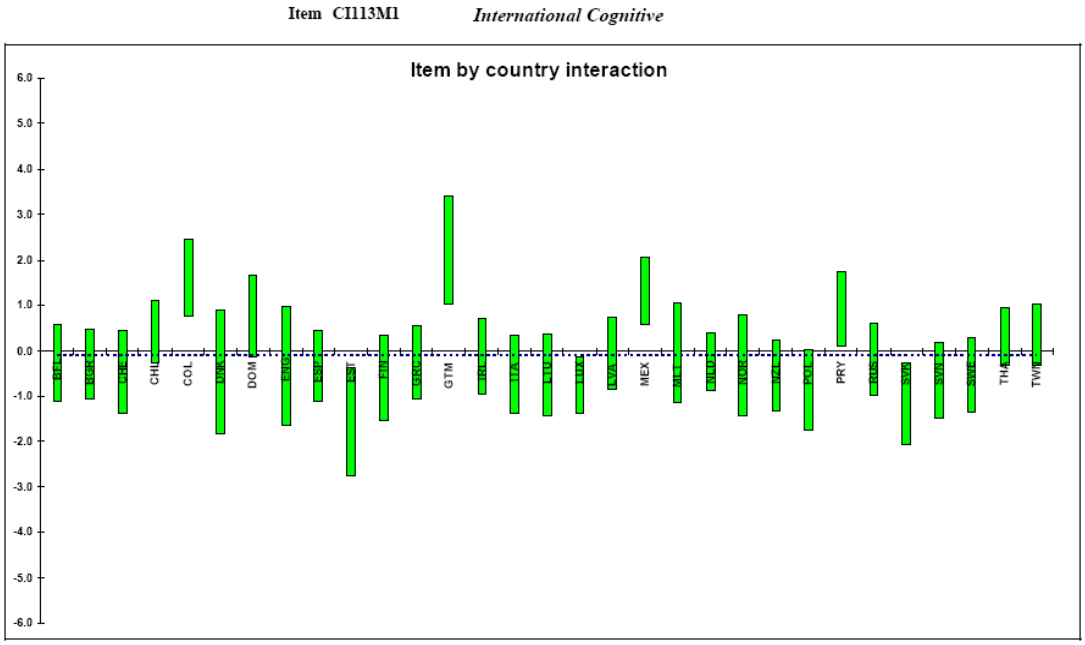
Figure 6 Graphical display of item-by-country interaction (example A)*



* ACER ConQuest estimates.

For each item used in the cognitive test there is one bar chart where the bars indicate the estimated item parameters for each participating country with their respective confidence intervals[5]. The dotted 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 6 shows an item in which there is relatively little item by country interaction. Example B in Figure 7 is for an item for which there is some evidence of systematic interaction by geographical region. In this example the results indicate that the item was relatively harder in Latin American countries.

Figure 7: Graphical display of item-by country interaction (example B)*



* ACER ConQuest estimates.

Generally, the ICCS test items showed only few indications of item-by-country interactions. Information about occurrence of cross-national DIF was used during the selection of items for the main survey test instrument.

Questionnaire Item Analysis with the Rasch Model

Three types of questionnaire items can be distinguished: (1) Items that are reported as single items (e.g. gender of students, teachers or principals), (2) items that are designed to be used for computing indices via an arithmetic transformation (e.g. student-teacher ratios from school questionnaire data) or (3) items that are used to derive scales (e.g. Likert-type items with agreement categories).

As with cognitive test items, percentages of categories and missing values were reviewed to ensure that selected main survey were not overly skewed and had sufficiently high percentages of valid responses. Student background questions were also analysed with regard to their relationship with indicators of civic knowledge and engagement.

[5] The confidence intervals for item parameters were adjusted for design effects (due to the cluster sample design) and multiple comparisons (for 31 countries).

The large majority of ICCS field trial items were used for scaling purposes and field trial data were used to undertake the following types of analysis:

- Exploratory factor analysis (EFA)
- "Classical" item and scale statistics
- Confirmatory factor analysis (CFA)
- Rasch modelling

In addition to comparative CFA and multiple-group analyses (see Schulz 2009), Rasch modelling with the Partial Credit Model (PCM) provided a useful tool for reviewing measurement equivalence of questionnaire items across national sub-samples. To this end models with country interaction effects (see equation 3) were estimated that provided indicators of the degree of parameter invariance across national sub-samples.

Table 1 shows a set of eight items that was trialled and designed to measure students' attitudes towards their own country covering both students' symbolic patriotism and uncritical patriotism. Four of these items had already been included in the IEA CIVED study in 1999 (see Schulz, 2004, p. 106f). Exploratory factor analyses showed that item IS2H02D (preferring to live in another country) had only weak factor loadings and was discarded from further scaling analyses.

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.

Table 2 shows the results from the international calibration using the pooled field trial sample and summaries of item-by-country interactions. The degree of parameter variation across countries was summarised to provide information about the degree of measurement equivalence. In the ICCS field trial analysis, the median of absolute values for item-by-country interaction effect was taken as an indicator of parameter

invariance for each item. In addition, the minimum and maximum effects were displayed to demonstrate the range of deviations across countries.

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

		Calibration results		Item-by-country interaction		
Item	Content	Parameter	Fit	Median of absolute values	Minimum	Maximum
IS2H02A	Flag important	-0.280	0.98	0.59	-1.003	1.183
IS2H02B	Great respect for country	-0.652	0.83	0.26	-1.006	0.681
IS2H02C	Proud of achievement	-0.586	0.86	0.11	-0.914	0.752
IS2H02E	Proud to live in country	-0.483	0.84	0.11	-0.416	0.439
IS2H02F	Support country always	0.594	1.25	0.27	-0.741	0.961
IS2H02G	Better country than others	0.350	1.10	0.48	-0.895	1.154
IS2H02H	World better place	1.056	1.22	0.25	-0.578	0.934

Note: ACER ConQuest estimates.

The results suggest that in particular items IS2H02F and IS2H02H do not fit well with the overall scale. Both items are less discriminating than others and also have somewhat lower item-total correlations. The summary of country DIF indicates that some of the items show higher median values of item-by-country interaction, in particular the item IS2H02A and IS2H02G.

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

ISO code for Country	Flag important IS2H02A	Great respect for country IS2H02B	Proud of country's achievement IS2H02C	Proud to live in country IS2H02E	Support country always IS2H02F	Better country than others IS2H02G	World better place IS2H02H
BFL	**1.18**	0.09	0.13	0.03	***-0.64***	***-0.38***	***-0.41***
BGR	***-0.58***	***-0.48***	-0.07	0.09	-0.16	**0.73**	**0.45**
CHE	**0.64**	0.19	0.10	-0.02	-0.07	***-0.74***	-0.10
CHL	***-0.39***	***-0.32***	-0.02	-0.01	0.27	0.22	0.25
COL	***-0.57***	***-0.44***	0.05	***-0.33***	**0.69**	**0.50**	0.10
DNK	**0.77**	**0.48**	-0.12	0.16	***-0.44***	***-0.62***	-0.24
DOM	***-0.89***	0.13	0.01	0.05	**0.96**	0.15	***-0.40***
ENG	**0.78**	0.13	-0.16	***-0.34***	0.18	***-0.32***	-0.28
ESP	0.18	0.15	0.17	***-0.42***	0.17	-0.15	-0.09
EST	0.02	-0.03	-0.24	0.16	-0.27	0.02	**0.34**
FIN	0.10	**0.42**	0.11	0.05	-0.21	***-0.72***	0.25
GRC	***-0.93***	***-0.47***	-0.03	0.10	-0.06	**0.49**	**0.90**
GTM	***-0.36***	-0.07	-0.15	-0.06	-0.19	**0.49**	**0.33**
IRL	0.10	***-0.30***	***-0.43***	***-0.38***	**0.40**	0.23	**0.39**
ITA	***-0.71***	***-0.30***	**0.75**	-0.18	0.09	0.13	0.21
LTU	***-0.76***	***-1.01***	***-0.91***	-0.11	**0.70**	**1.15**	**0.93**
LUX	**0.69**	**0.68**	0.18	0.12	***-0.50***	***-0.76***	***-0.41***
LVA	***-0.36***	0.25	0.23	**0.37**	***-0.36***	0.12	-0.25
MEX	***-0.70***	-0.26	-0.03	-0.09	**0.49**	**0.48**	0.12
MLT	-0.17	0.11	-0.11	0.14	-0.03	0.04	0.04
NLD	**1.08**	0.27	0.14	0.14	***-0.35***	***-0.70***	***-0.58***
NOR	**0.74**	0.12	-0.01	-0.09	0.03	***-0.90***	0.10
NZL	**0.83**	0.11	-0.03	-0.15	-0.06	***-0.32***	***-0.39***
POL	***-0.59***	-0.25	0.25	**0.44**	***-0.74***	**0.68**	0.20
PRY	***-1.00***	***-0.62***	0.06	0.14	**0.94**	**0.48**	0.00
RUS	0.21	0.08	0.08	-0.10	-0.14	0.09	-0.22
SVK	0.09	0.02	-0.13	-0.03	***-0.70***	**0.78**	-0.05
SVN	0.15	**0.41**	0.09	0.02	-0.11	-0.22	***-0.34***
SWE	**0.82**	**0.59**	-0.05	-0.09	***-0.46***	***-0.60***	-0.21
THA	-0.20	-0.21	-0.05	0.04	**0.62**	-0.04	-0.17
TWN	-0.19	**0.53**	0.18	**0.35**	-0.06	***-0.34***	***-0.48***

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

Table 3 shows the estimated item-by-country interaction terms for the item location parameter. Positive signs indicate that an item was relatively harder to agree with than in others, negative signs that it was more frequently endorsed compared than internationally.

Item IS2H02A ("importance of flag") has the highest median country DIF and reviewing the country-level results reveals that students in many developed countries (Northern and Central Europe, New Zealand) find it relatively harder to agree with this item whereas students from Latin American or Southern European countries tend to agree more readily with this item. In other words, the item location parameters would be quite different when calibrating in Latin America compared to those from a

calibration in Scandinavian countries. Interestingly, a similar pattern emerges for item IS2H02B ("great respect for country"), the other item reflecting symbolic patriotism. However, this item appears to have less item-by-country interaction.

For item IS2H02G ("own country better to live in than others") there is also considerable item-by-country interaction. This items tends be relatively easier to endorse in countries developed countries particularly in Northern Europe whereas students find it harder to agree with in Latin American and Eastern European countries. This means that considerably lower item location parameters would be estimated in wealthier countries compared to those that would results from calibrations in poorer countries. Item IS2H02H ("word better place if all were like country's citizens") shows a somehow similar pattern but generally less item-by-country interaction.

In summary, the results for the items measuring students' attitude towards their own country show that there is a noticeable lack of measurement equivalence. It is interesting to note that those items related to symbolic patriotism ("importance of flag", "great respect for country") are relatively easier to agree with in Latin American and Mediterranean countries and relatively harder to be endorsed in Northern European countries. Items that may be more affected by objective living conditions ("better country to live in than others", "world better place if all were like country's citizens") on the other hand appear to be relatively harder to agree with in poorer countries.

Information like the one presented in this example was used for item selection for the main survey in conjunction with other indicators and generally preference was given to items that had less country DIF. However, differential behaviour of items can also be quite informative and it should not automatically lead to the exclusion of items. In this particular case, the items IS2H02A, IS2H02B and IS2H02G that had considerable item-by-country interaction were retained for the main survey with slightly modified set of items, in particular because the first two items had also been used in the previous IEA CIVED study.

Conclusion and Discussion

The Rasch item parameters provide a rich set of data with which to investigate item-by-country interaction. In the cognitive test, it was possible to consider both the overall item performance across all countries, which impacts on the integrity of the international scale, and the individual item performance within countries against the international scale, which influences individual country statistics. Ultimately, in selecting items for the final instruments, a balance needs to be struck between a level of item variability that will not compromise the integrity of the instrument measures and the necessity to have sufficient items to cover the breadth of the Assessment Framework. Fortunately, overall the ICCS field trial data and analysis revealed relatively few items that showed unacceptably high levels of item-by-country interaction.

It should be noted, that the focus of decisions about the degree of item-by country interaction has been on the relative scaled item difficulties between individual countries and the international scale. Data were provided at a country level on the broader range of item function (such as fit to the model of individual items compared to the international scale) but typically, unless an individual item was functioning so

poorly as to compromise its measurement integrity. Information about item-by-country interaction was part of the evidence regarding the item functioning but did not automatically determine the item selection process.

Further investigation of the effects of other parameters on the measurement invariance of the ICCS test instrument could provide an informative and valuable extension to the existing ICCS cognitive test item-by-country interaction analysis. For example, hierarchical cluster analysis of item-by-country interaction parameters might provide further information to what extent variation is influenced by cultural or language background of participating countries.

After main survey data have been collected, questionnaire items in ICCS will be scaled using IRT. Therefore, it is important to assess the appropriateness of the assumption of using the scaling model with internationally determined item parameters across countries. ACER ConQuest allows estimating item-by-country interaction parameters directly to review measurement equivalence for questionnaire scales.

The example shown in this paper illustrates 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. Often item parameter variation is informative and should not only be viewed as "undesirable measurement bias". Consequently, it is not recommended to apply any automatic exclusion rules based on this type of analysis when selecting items for the main survey. Occurrence of larger item-by-country interaction, however, may become one criterion for selection when reviewing individual items. In cases where all items used for construct measurement show higher levels of parameter invariance, researcher might decide to rather omit this instrument or opt for an alternative approach.

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, there are some 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, for example, Heine, Lehman, Peng, & Greenholtz, 2002) and the review of parameter invariance could be extended to the functioning of step parameters (see, for example, Walker, 2007).

The analyses of item-by-country interactions undertaken with ICCS field trial data show a noticeable but limited amount of country DIF. 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, data on parameter invariance can only be interpreted as relative measures.

In the case of ICCS, information on item-by-country interaction helped to select items for the main survey that showed lower levels of parameter variation across countries. Some of the variation is probably inevitable because test and questionnaire instruments are translated into numerous languages and administered in many different contexts around the world. In addition, parameter invariance can also show interesting differences depending on cultural or educational contexts and the question arises to what extent it is really desirable to only use items that measure exactly in the same way across countries.

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 reduces comparisons to those issues that are relatively uniform across countries. An important question that 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.

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).

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

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

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

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

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

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

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

Schulz, W. (2004). Scaling Procedures for Likert-type Items on Students' Concepts, Attitudes and Actions. In W. Schulz & H. Sibberns (Eds.) *IEA Civic Education Study. Technical Report (*pp. 93-126). Amsterdam: IEA.

Schulz, W. (2009). Questionnaire Construct Validation in the International Civic and Citizenship Education Study. In *IERI Monograph Series* Volume 2, 85-107.

Schulz, W. & Sibberns, H. (2004) (Eds.) *IEA Civic Education Study. Technical Report*. Amsterdam: IEA.

Schulz, W., Fraillon, J., Ainley, J., Losito, B., & Kerr, D. (2008) *International civic and citizenship education study. Assessment framework.* Amsterdam: International Association for the Evaluation of Educational Achievement (IEA).

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

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

Walker. M. (2007). Ameliorating Culturally Based Extreme Item Tendencies to Attitude Items. *Journal of Applied Measurement*, 8(3), 267-278.

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