# Testing Parameter Invariance for Questionnaire Indices using Confirmatory Factor Analysis and Item Response Theory

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

Paper prepared for the Annual Meetings of the American Educational Research Association in San Francisco, 7-11 April 2005.

# Introduction

International studies like PISA, TIMSS or PIRLS use context student or school questionnaires to collect data on student family background, attitudes and learning context. Questionnaire constructs are typically measured using dichotomous or Likert-type items. Data from questionnaires are often used to explain variation in student performance or reported as learning outcomes.

Smith (2004) states regarding the challenges of developing questionnaire material in multicultural settings that the "very differences in language, culture and structure that make cross-national research so analytically valuable, seriously hinder achieving measurement equivalency" (p. 431). And Douglas and Nijssen (2003) warn that using "borrowed" scales that are developed in one specific national context cross-nationally may be "fraught with danger" when there is substantial variation in languages and cultures.

Whereas international studies tend to spend considerable efforts on ensuring measurement equivalence in international test instruments, the issue of equivalency of questionnaire data has not received quite the same attention. But given the importance of contextual information for the reporting and analysis of cross-national learning outcomes, the scaling of questionnaire items measuring family background, student attitudes or perceptions of learning context requires a thorough cross-country validation of the underlying constructs.

This paper describes ways of implementing tests of parameter invariance using both Confirmatory Factor Analysis (CFA) and Item Response Theory (IRT), examines results using these two different approaches to validate questionnaire constructs based on data from the OECD PISA study and discusses implications for cross-national or cross-cultural research.

# Cross-national Validation in International Survey Research

One of the most important goals of international studies in educational research is the comparison of learning outcomes across participating countries. In order to compare results it is necessary to collect data using *comparable measures*. Studies like TIMSS, CIVED, PIRLS or PISA invest considerable efforts in attempts to develop tests which are appropriately translated into the test languages, culturally unbiased and suitable for the diverse educational systems across participating countries. Typically, IRT (Item Response Theory) scaling methodology (see Hambleton, Swaminathan and Rogers, 1991) is used to review Differential Item Functioning (DIF) for countries and detect country-specific item misfit (see examples in Adams, 2002; Schulz and Sibberns, 2004).

Likewise, it is of great importance to achieve similar levels of comparability for measures derived from contextual questionnaires. Data collected from contextual questionnaires are often used to explain variation in student performance. However, many constructs measured in student questionnaire (for example self-related cognitions regarding areas of learning, classroom climate etc.) can often also be regarded as important learning outcomes.

In the OECD PISA, for example, study contextual data are collected through student and school questionnaires. Questionnaire items are treated in three different ways (see OECD, 2005, pp. 271-319):

- They are reported as single items (for example gender, grade).
- They are converted into "simple indices" through the arithmetical transformation or recoding of one or more items.
- They are scaled. Typically, Item Response Theory (IRT) is used as scaling methodology in order to obtain individual student scores (Weighted Likelihood Estimates).

Language differences can have a powerful effect on equivalence (or non-equivalence). Typically, source versions (in English or French) are translated into the language used in a country. In most international studies, reviews of national adaptations and thorough translation verifications are implemented in order to ensure a maximum of "linguistic equivalence" (see Grisay, 2002; Chrostowski and Malak, 2004). However, it is well known that even slight deviations in wording (sometimes necessary due to linguistic differences between source and target language) may lead to differences in item responses (see Mohler, Smith and Harkness, 1998; Harkness, Pennell and Schoua-Glusberg, 2004).

Another source of non-equivalence is the cultural diversity among participating countries in international studies. Cultural habits may have an influence on the degree to which respondents endorse certain item statements. In addition, differences between educational systems (with different instructional practices and policies) may impact on how questionnaire items are understood and interpreted. For example, statements indicating unfavourable learning context (disruptions at the beginning of each lesson) might be interpreted differently depending on instructional practices (see Schulz, 2003).

As pointed out by van de Veijver and Tanzer (1997), instruments might work properly but characteristics of cultural groups of respondents may introduce bias in measurement. Byrne (2003) distinguishes three different kind of bias related to cross-cultural research that may result from cultural differences between countries:

- *Construct bias* refers to cases where a construct may be meaningful in one country, but not another country.
- *Method bias* refers to cases where data are biased by differences in responses to the instruments caused by cultural traits (for example, respondents in some countries tend to systematically choose more extreme values in Likert-type scales than in other countries).
- *Item bias* refers to bias that occurs at the level of the individual item. Constructs might be well measured in general, but some items may exhibit differential item functioning due to cultural differences (for example, an item may be understood differently in culture and display different item characteristics compared to other countries).

Lack of equivalence in measures across participating countries in international studies can be due to different factors and it may become difficult to disentangle them. When using indicators to measure latent constructs in international studies, it is important to ensure that the variation which is measured between countries lies in the measured construct and that measured variation is not only due to the specific variation of unique indicators. If the specific variation of indicators (that is not due to the construct) is influenced by country differences, it will be problematic to make valid construct comparisons.

# Data and methods

The OECD PISA assesses the performance of 15-year-old students in the 30 OECD member countries and a growing number of non-OECD countries. The following instruments are regularly used to collect data in the OECD PISA study:

- Students are assessed with a 2-hour rotated test design that includes an extensive test on the major domain (2000: Reading, 2003: Mathematics, 2006: Science) and smaller subtests for minor domains (alternating Mathematics, Reading, Science and Problem Solving as an additional area of assessment in 2003).
- The student questionnaire includes questions on student characteristics, home background, educational career, school/classroom climate, learning behaviour and self-related cognitions in the area of the major domain (reading, mathematics or science).
- The school questionnaire collects data on the school characteristics and learning environment and is addressed to the school principal.

The data presented in this paper were collected in the field trial for the PISA 2006 study, which was carried out in all participating countries between March and September 2005 using convenience samples (roughly representing all major school types and study programmes) of typically 1200 students.

In order to avoid complexity in the presentation of data and providing some kind of "post-hoc experimental design" for the analyses, only a selection of 16 country samples from eight different groups of countries was used. The selected countries were chosen so that there were two countries with similar or identical languages and cultural background from each group.

Using such a design allows the assessment of whether differences in parameters are rather influenced by country-specific factors or the language and cultural background common for both countries in each group. Table 1 shows the countries and groups selected for the analyses.[1]

**Table 1 Groups and countries in analyses**

| Country | Group | Sample size* |
|---|---|---|
| Tunisia | Arabic-speaking (ARA) | 1224 |
| Jordan | Arabic-speaking (ARA) | 1462 |
| Netherlands | Dutch-speaking (BEN) | 1262 |
| Belgium (Flemish) | Dutch-speaking (BEN) | 1488 |
| Chinese Taipeh | Chinese-speaking (CHI) | 2121 |
| Hong Kong | Chinese-speaking (CHI) | 1207 |
| Slovak Republic | Slavic-speaking (CZS) | 1864 |
| Czech Republic | Slavic-speaking (CZS) | 1365 |
| New Zealand | English-speaking (ENG) | 1165 |
| Australia | English-speaking (ENG) | 1992 |
| Germany | German-speaking (GER) | 5642 |
| Austria | German-speaking (GER) | 1955 |
| Norway | Scandinavian (SCA) | 1188 |
| Sweden | Scandinavian (SCA) | 1214 |
| Chile | Spanish-speaking (SPA) | 2648 |
| Uruguay | Spanish-speaking (SPA) | 1312 |

* Due to rotated questionnaire design only about half the sample size was available for each of the analyses.

Some groups are obviously more homogenous than other. Whereas most country groups have the same language and of similar cultural background, others are more heterogeneous: The Chinese-speaking group with Chinese Taipeh and Hong Kong could be considered as rather heterogenous given the cultural and historical background of both entities. In the member countries of the Scandinavian group (Norway and Sweden) different languages are spoken, however, the languages are very similar. The same can be said about the group of Slavic-speaking countries.

Two methodological approaches to construct validation have been applied within the context of international studies (see examples in Schulz, 2002; Schulz, 2004; OECD, 2005, pp. 271-319):

- Confirmatory Factor Analysis (CFA) based on mean and covariance structures are used to test the measurement equivalence and help to detect possible socio-cultural differences on the constructs.
- Item Response Theory (IRT) is used for scaling items using a logistic function and obtaining measures of the latent construct. In addition, it is also useful for assessing item dimensionality and differential item functioning across countries.

[1] As the data were obtained in a field trial and are not officially released, country names will not be displayed in any tables but just numbered within each group of countries.

Both CFA and IRT scaling methodology provide tools for reviewing the cross-cultural validity of questionnaire constructs. Little (1997) proposes to extend the use of CFA to multiple-group analysis of mean and covariance structures (MACS) for testing the comparability of measurement equivalence of psychological constructs and detecting possible socio-cultural variation of factor loadings and intercept parameters.

IRT methodology has also been used to detect non-equivalence across countries, in particular with regard to different response patterns when using Likert-type items (see for example Andrich and Luo, 2003; Maris and Maris, 2002). Research comparing both methodological approaches (Wilson, 1994) has suggested that IRT provides a more rigid test of parameter invariance across countries than covariance-based methods and has a higher likelihood to lead to the rejection of the hypothesis of measurement equivalence.

# Confirmatory Factor Analysis

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

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

where $\Lambda_x$ is a q x k matrix of factor loadings, $\xi$ denotes the latent variable(s) and $\delta$ is a q x 1 vector of unique error variances.

The expected covariance matrix is fitted according to the theoretical factor structure. With continuous variables, Maximum Likelihood (ML) estimation provides model estimates trying to minimise the differences between the expected (Σ) and the observed covariance matrix (S). However, Maximum Likelihood (ML) and Generalised Least Square (GLS) estimation both require normal distribution and continuous variables.

For non-normal, ordinal variables Jöreskog and Sörbom (1993) recommend to use *Weighted Least Square Estimation* (WLS) with polychoric correlation matrices and corresponding asymptotic covariance weight matrices. However, a common problem with estimation methods that explicitly address non-normal and categorical manifest variables is the reliance on very large sample sizes or unrealistically small models. Recent developments in the WLS based estimation of structural model parameters under non-normality now do not require such large sample sizes. Based on work by Satorra (1992), Muthén and his colleagues (for example in O. Muthén, du Toit, and Spisic, 1997) developed a mean-adjusted WLS estimator (WLSM) and mean- and variance- adjusted WLS estimator (WLSMV). Both estimation methods are available in the Mplus software program and the WLSMV estimator was used for the CFA presented in this paper.

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

In international studies, the parameters may vary across country and it may not be appropriate to assume the same factor structure for each population. One way of looking at invariance of factor structures is to use separate CFA within countries and review model fit within each population across countries.[2] Though this will provide insights into the validity of assuming the same factor structure, it does not enable us to test the invariance of model parameters.

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

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

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

(3) $H_\Lambda = \Lambda_1 = \Lambda_1 = \Lambda_2 = ... = \Lambda_g$

Hypothesis testing using tests of significance tends to be problematic, in particular with data form large samples where even smaller differences appear to be significant. Therefore, a modelling approach looking at relative changes in model fit is preferable. This can be done by setting placing different equality constraints on parameters in multiple-group models and comparing model fit indices across different multiple-group models with increasing constraints starting with a totally unconstrained model.

Different types of constraints can be used in order to review the invariance of model parameters. Once the invariance of factor structure and factor loadings has been confirmed, further constraints might be placed on factor variances and covariances. However, observing different factor variances and covariances may be a finding rather than an indication of factor structure invariance. For example, it might be unrealistic to expect that two constructs related to learning and teaching have the same correlation regardless of the educational policies and practices implemented in different countries. Therefore, it seems reasonable to assume invariance of factor loadings as a sufficient condition for construct validity across countries.

There are no clear criteria for judging model fit in SEM techniques: Chi-square test statistic for the null hypothesis of *=S become rather poor fit measures with larger sample sizes because even small differences between matrices are given as significant deviations. Therefore, recent practice gives emphasis to alternative fit indices like the Root Mean Square Error of Approximation (RMSEA), which measures the "discrepancy per degree of freedom for the model" (Browne and Cudeck, 1993: 144). A value of .05 and less indicates a close fit, values of .08 and more indicate a reasonable error of approximation and values greater than 1.0 indicate poor model fit.

In addition, model fit can be assessed using the Comparative Fit Index (CFI) and the Tucker-Lewis Index (TLI) (also known as the Non-normed Fit Index, NNFI) which are less dependent on sample size and correct for model complexity (see Bollen and Long, 1993). CFI and TLI should have values close to 1 in order to indicate good model fit.

[2] Please note that it is also recommended to use Exploratory Factor Analysis (EFA) to analyse whether same factor structures can be observed for each country. This may become a particularly useful approach when assumptions about item dimensionality are not clearly defined.

In the multiple-group analyses presented in this paper four different models will be tested. As chi square based tests of statistical significance tend to be problematic with larger sample size, the results should be judged according to "relative model fit" of models with different degrees of constraints.

**Table 2 Description of multiple-group models in analysis**

| | Constraints |
|---|---|
| Model 1 | Unconstrained model |
| Model 2 | Constraints on factor loadings within groups of countries |
| Model 3 | Constraints on factor loadings across countries |
| Model 4 | Constraints on factor loadings, factor variances and covariances |

Table 2 shows four different multiple-group models: Starting from a totally unconstrained model, in a first step factor loadings are constrained within groups of countries. In a second step factor loadings are constrained across all countries and in a third step additional constraints are placed on factor variances and covariances.[3]

## Item Response Theory

PISA questionnaire items are typically scaled using IRT (Item Response Theory) scaling methodology (see Hambleton, Swaminathan and Rogers, 1991). With the One-Parameter (Rasch) model (Rasch, 1960) for dichotomous items, the probability of selecting category 1 instead of 0 is modelled as

$$(4) \qquad P_i(\theta) = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)},$$

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

For items with more than two (k) categories (as for example with Likert-type items) this model can be generalised to the *Partial Credit Model* (Masters and Wright, 1997)[4], which is defined as

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

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

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

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

Item fit can be assessed using the weighted mean-square statistic (infit), a residual-based fit statistic. Weighted *infit* statistics can be computed both for item and step parameters.

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

Equation (5) shows that the part of the model related to the item consists of the item parameter $\delta_i$ for item i and the step parameter $\tau_{ij}$ for step j of item i. When using the scaling software ACER ConQuest (Wu, Adams and Wilson, 1997), the model term $\delta_i + \tau_{ij}$ is described with the statement ITEM+ITEM*STEP. For the purpose of the analysis of parameter equivalence, additional parameters for country group (GRP) or country effects (CNT) can be added to this model. Table 3 shows the models used to review item parameter invariance in this paper.

**Table 3 Description of IRT models used to review parameter invariance**

| Model | Description | ConQuest Model statement |
|---|---|---|
| 1 | Items and steps constrained | ITEM+ITEM*STEP |
| 2 | Items unconstrained by group | ITEM-GRP+ITEM*GRP+ITEM*STEP |
| 3 | Items unconstrained by country | ITEM-CNT+ITEM*CNT+ITEM*STEP |
| 4 | Items & steps unconstrained by country | ITEM-CNT+ITEM*CNT+CNT*ITEM*STEP |

The first model is the one which is used when items are scaled with item parameters obtained from a calibration of a pooled international sample. The second model assumes that there are effects of "groups of countries" on the item location parameters. This is modelled as an interaction term (ITEM*GRP) and in order to get the correct estimates it is necessary to add another parameter for the country group effect on item responses (GRP).[5]

The third model replaces the country group effect (GRP) with the effect of individual countries (CNT). This is done to review whether parameter invariance is more influenced by country-specific factors than by language and culture common to more than one country. The fourth model goes one step further by adding a country interaction with step parameters (CNT*ITEM*STEP) so that for each item in each country separate step parameters are estimated.

Comparisons of these different scaling models can be guided by differences in the deviance statistics obtained from each calibration. Chi square tests can be obtained by taking the differences in parameters into account. However, when working with larger sample sizes chi square tests even minor differences may appear as significant.

Furthermore, models with country group or county interaction effects provide estimates of the degree of parameter invariance across countries or groups of countries. The degree of parameter variation across countries can be displayed graphically to inform about the degree of measurement equivalence. Again, tests of significance will most likely be significant in view of the larger sample sizes used in

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

international studies and "rules of thumb" need to be developed in order to judge the extent of "tolerable" parameter invariance.

# Empirical Example: PISA 2006 Field Trial Questionnaire Data

The student questionnaire used in the PISA 2006 Field Trial covered a wide range of issues including student home background, student self-beliefs in science (as the major domain in the PISA test), motivational factors to learn science, teaching and learning of science, parenting styles and environmental issues.

In order to trial a wider range of questionnaire material, four rotated questionnaire forms were used. Not all of the questionnaire material was retained after the field trial, some items or constructs were ruled out on empirical grounds (item misfit, lack of dimensionality), others were discarded on conceptual grounds (priorities had to be assigned as the main study questionnaire is limited to 30 minutes of assessment time). For the analyses in this paper, two different sets of student questionnaire constructs were chosen to illustrate the analysis of parameter invariance.

**Table 4 Items and constructs on self-concepts in science, mathematics and reading***

| | | | |
|---|---|---|---|
| **Q20** | | **How much do you agree with the following statements?** | |
| ST20Q01 | a | Learning advanced <science topics> would be easy for me | SCSCIE |
| ST20Q02 | b | I can usually give good answers to <test questions> on <science topics> | SCSCIE |
| ST20Q03 | c | I learn <science topics> quickly | SCSCIE |
| ST20Q04 | d | <Science topics> are easy for me | SCSCIE |
| ST20Q05 | e | When I am being taught <science>, I can understand the concepts very well | SCSCIE |
| ST20Q06 | f | I can easily understand new ideas in <science> | SCSCIE |
| **Q21** | | **How much do you agree with the following statements?** | |
| ST21Q01 | a | Learning advanced mathematics topics would be easy for me | SCMATH |
| ST21Q02 | b | I can usually give good answers to <test questions> on mathematics topics | SCMATH |
| ST21Q03 | c | I learn mathematics topics quickly | SCMATH |
| ST21Q04 | d | Mathematics topics are easy for me | SCMATH |
| ST21Q05 | e | When I am being taught mathematics, I can understand the concepts very well | SCMATH |
| ST21Q06 | f | I can easily understand new ideas in mathematics | SCMATH |
| **Q22** | | **How much do you agree with the following statements?** | |
| ST22Q01 | a | Learning advanced <test language> topics would be easy for me | SCREAD |
| ST22Q02 | b | I can usually give good answers to <test questions> on <test language> topics | SCREAD |
| ST22Q03 | c | I learn <test language> topics quickly | SCREAD |
| ST22Q04 | d | <Test language> topics are easy for me | SCREAD |
| ST22Q05 | e | When I am being taught <test language>, I can understand the concepts very well | SCREAD |
| ST22Q06 | f | I can easily understand new ideas in <test language> | SCREAD |

* Expressions in <> were adapted to the national context of each country.

One set deals with students' self-concept in science (SCSCIE), mathematics (SCMATH) and reading (SCREAD). Due to lower priorities assigned to the self-concepts in mathematics and reading, only the construct of science self-concept was retained for inclusion in the main study questionnaire. The wording of each set of items is identical except for the subject area (see Table 4).

## Table 5 Items and constructs on science teaching practices

| ST36 | | When learning <science topics> at school, how often do the following activities occur? | |
|---|---|---|---|
| ST36Q01 | a | Students spend time in the laboratory doing practical experiments | HANDSON |
| ST36Q09 | i | Students are asked to draw conclusions from an experiment they have conducted | HANDSON |
| ST36Q11 | k | Experiments are done by the teacher as demonstrations | HANDSON |
| ST36Q14 | n | Students do experiments by following the instructions of the teacher | HANDSON |
| ST36Q03 | c | Students are required to design how a <science> question could be investigated in the laboratory | SCINVEST |
| ST36Q05 | e | Students are allowed to design their own experiments | SCINVEST |
| ST36Q12 | l | The teacher gives students the chance to choose their own investigations | SCINVEST |
| ST36Q16 | p | Students are asked to do an investigation to test out their ideas | SCINVEST |
| ST36Q04 | d | The teacher starts a topic with an everyday example | SCMODEL |
| ST36Q07 | g | The students are asked to apply a <science> topic to everyday problems | SCMODEL |
| ST36Q13 | m | The teacher uses science to help students understand the world outside school | SCMODEL |
| ST36Q15 | o | The teacher clearly explains the relevance of <science> concepts to our lives | SCMODEL |
| ST36Q17 | q | The teacher uses examples of technological application to show how <science> is relevant to society | SCMODEL |

* Expressions in <> were adapted to the national context of each country.

The other set includes items measuring students' perception of teaching methods: Hands-on teaching (HANDSON), student investigations (SCINVEST) and use of models or applications (SCMODEL). Table 5 shows the wording of the items and the constructs it measures.[6] As the question stem refers to the learning of science at school, it should be noted that in many countries students had to make judgements about what happens in lessons taken in different science subjects (like biology, chemistry or physics).

When reviewing field trial outcomes in PISA, the following statistics are routinely reported for questionnaire items: International frequencies (percentage valid and missing), national scale reliabilities, item-total correlations and correlations with student performance. Item-total correlations are in particular useful for detecting possible errors in translation, for example, a negative correlation of an item with the total scale in a particular country would be discussed with its national centre staff asking them to re-check translation and/or adaptation of this "dodgy" item. An overall CFA for each set of items is estimated in order to review whether the assumption of unidimensionality holds in the pooled international sample.

In the analyses undertaken for this paper, prior to the multiple-group models, separate models for each country data set were estimated in order to review whether the factor structure holds also within each country. The software program MPLUS was used for estimating all separate group models and multiple-group models.[7]

Table 6 shows the results for the separate three-factor models for self-concept items and the results for multiple-group models. The fit measures indicate a reasonable fit for the three-factor structure in most countries. In three countries, however, the model fit is rather poor (> .1). Poor model fit does not occur in a whole group of countries which indicates that it is not clearly related to common language or cultural backgrounds.

[6] Initially, different dimensions had been anticipated but were not confirmed by the field trial analysis. The dimensionality for the sets used in this example was determined using exploratory factor analysis and in consultation with experts in science teaching.

[7] Poly-choric correlations and the mean- and variance- adjusted WLS estimator (WLSMV) were used for model estimation.

When looking at the estimated latent correlations between the three constructs, it appears that there is substantial variation across countries. Whereas science and mathematics self-concept are positively correlated in most countries, correlations between science and reading as well as between reading and mathematics range both from negative to positive correlations. However, this does not necessarily indicate any lack of measurement equivalence as it is likely to be an empirical finding due to different instructional practices and learning contexts. Whereas in both Scandinavian countries there is a (modest) positive correlation between self-concept in reading and mathematics, in both Dutch-speaking countries the correlation is negative.

**Table 6  CFA results for self-concept items**

| | **Model fit** | | | **Latent Correlations of ...** | | |
|---|---|---|---|---|---|---|
| **Country** | **CFI** | **TLI** | **RMSEA** | **SCSCIE with SCMATH** | **SCSCIE with SCREAD** | **SCMATH with SCREAD** |
| | | | **Separate group models** | | | |
| **ARA1** | 0.98 | 0.99 | 0.066 | 0.34 | 0.05 | -0.15 |
| **ARA2** | 0.97 | 0.99 | 0.072 | 0.64 | 0.23 | -0.02 |
| **BEN1** | 0.98 | 0.99 | 0.083 | 0.49 | -0.12 | -0.19 |
| **BEN2** | 0.97 | 0.99 | 0.085 | 0.42 | 0.09 | -0.15 |
| **CHI1** | 0.98 | 0.99 | 0.119 | 0.51 | -0.07 | -0.28 |
| **CHI2** | 0.98 | 0.99 | 0.080 | 0.57 | 0.02 | -0.04 |
| **CZS1** | 0.97 | 0.99 | 0.095 | 0.32 | 0.35 | 0.13 |
| **CZS2** | 0.97 | 0.99 | 0.090 | 0.26 | 0.14 | -0.03 |
| **ENG1** | 0.98 | 0.99 | 0.088 | 0.41 | 0.26 | 0.08 |
| **ENG2** | 0.98 | 1.00 | 0.073 | 0.39 | 0.22 | 0.07 |
| **GER1** | 0.99 | 1.00 | 0.058 | 0.37 | 0.22 | -0.09 |
| **GER2** | 0.97 | 0.99 | 0.100 | 0.30 | 0.16 | -0.06 |
| **SCA1** | 0.95 | 0.98 | 0.152 | 0.64 | 0.22 | 0.15 |
| **SCA2** | 0.99 | 1.00 | 0.071 | 0.54 | 0.33 | 0.18 |
| **SPA1** | 0.98 | 0.99 | 0.073 | 0.43 | 0.25 | -0.03 |
| **SPA2** | 0.99 | 1.00 | 0.066 | 0.19 | 0.18 | -0.11 |
| ***Median*** | ***0.98*** | ***0.99*** | ***0.080*** | ***0.41*** | ***0.20*** | ***-0.04*** |
| | | | **Multiple-group models*** | | | |
| **Model 1** | 0.97 | 0.99 | 0.088 | | | |
| **Model 2** | 0.97 | 0.99 | 0.088 | | | |
| **Model 3** | 0.96 | 0.99 | 0.091 | | | |
| **Model 4** | 0.99 | 0.99 | 0.103 | 0.43 | 0.18 | -0.04 |

* Model 1 is unconstrained; Model 2 places constraints on factor loadings within groups of countries, Model 3 on loadings across all countries and Model 4 on factor loadings, variances and covariances.

The multiple-group model without constraints has an acceptable, albeit poor model fit. When placing constraints within groups of countries (that is estimating the same factor loadings within each group), the overall model fit is not very different from the one for the unconstrained model. Placing constraints on factor loadings across all datasets (that is estimating the same factor loading for all countries) leads to minor change in the fit indices but the overall fit still appears to be acceptable. The RMSEA for the multiple-group model with (additional) constraints on factor variances and covariances indicates a poor model fit (though TLI and CFI still indicate a reasonable fit). This is not surprising as it could already be observed that the relationships between constructs vary across countries and that assuming equality of parameters is not appropriate for this set of constructs.

Table 7 shows the results for separate group and multiple-group models for science teaching method items. The three-factor structure has a reasonable fit in most countries but model fit is poor in both Chinese-speaking and both Scandinavian countries as well as in one of the Spanish-speaking countries. So with this set of constructs, there is some indication that lack of fit is associated with characteristics of country groups.

**Table 7 CFA results for science teaching items**

| | **Model fit** | | | **Latent Correlations of...** | | |
|---|---|---|---|---|---|---|
| Country | **CFI** | **TLI** | **RMSEA** | **HANDSON with SCINVEST** | **HANDSON with SCMODEL** | **SCINVEST with SCMODEL** |
| | | | **Separate group models** | | | |
| **ARA1** | 0.84 | 0.88 | 0.070 | 0.91 | 0.78 | 0.80 |
| **ARA2** | 0.92 | 0.96 | 0.078 | 0.84 | 0.83 | 0.92 |
| **BEN1** | 0.94 | 0.98 | 0.066 | 0.79 | 0.68 | 0.57 |
| **BEN2** | 0.93 | 0.97 | 0.080 | 0.72 | 0.78 | 0.59 |
| **CHI1** | 0.88 | 0.95 | 0.106 | 0.74 | 0.66 | 0.79 |
| **CHI2** | 0.86 | 0.95 | 0.132 | 0.91 | 0.82 | 0.76 |
| **CZS1** | 0.90 | 0.95 | 0.075 | 0.84 | 0.54 | 0.55 |
| **CZS2** | 0.92 | 0.96 | 0.074 | 0.82 | 0.57 | 0.61 |
| **ENG1** | 0.88 | 0.95 | 0.097 | 0.85 | 0.74 | 0.68 |
| **ENG2** | 0.89 | 0.95 | 0.093 | 0.82 | 0.73 | 0.65 |
| **GER1** | 0.90 | 0.95 | 0.081 | 0.79 | 0.57 | 0.64 |
| **GER2** | 0.94 | 0.97 | 0.066 | 0.74 | 0.65 | 0.79 |
| **SCA1** | 0.77 | 0.91 | 0.122 | 0.80 | 0.61 | 0.74 |
| **SCA2** | 0.84 | 0.93 | 0.111 | 0.88 | 0.66 | 0.69 |
| **SPA1** | 0.89 | 0.95 | 0.093 | 0.92 | 0.79 | 0.92 |
| **SPA2** | 0.85 | 0.91 | 0.108 | 0.80 | 0.71 | 0.74 |
| ***Median*** | ***0.89*** | ***0.95*** | ***0.087*** | ***0.82*** | ***0.69*** | ***0.72*** |
| | | | **Multiple-group models** | | | |
| **Model 1** | 0.85 | 0.94 | 0.096 | | | |
| **Model 2** | 0.84 | 0.94 | 0.097 | | | |
| **Model 3** | 0.80 | 0.93 | 0.106 | | | |
| **Model 4** | 0.86 | 0.93 | 0.105 | 0.82 | 0.68 | 0.71 |

* Model 1 is unconstrained; Model 2 places constraints on factor loadings within groups of countries, Model 3 on loadings across all countries and Model 4 on factor loadings, variances and covariances.

The three constructs are highly correlated across all countries, hands-on activities and student investigations are correlated around 0.8 in most countries. Correlations between the three constructs are particularly high in the two Arabic-speaking countries. However, relationships between the three construct appear to be relatively uniform across the analysed datasets.

The multiple-group mode with unconstrained parameters shows an acceptable fit (when looking at RMSEA and CFI). When constraining factor loadings to be equal within each country group, the fit indices do not change much. This is in line with the observation from the separate group model estimations that lack of equivalence might be associated with the country group rather than with the individual country.

Constraining factor loadings to be equal across all countries leads to very poor model fit indices (Model 3) which indicates a certain lack of parameter invariance for this set of constructs. Given the relative homogeneity of latent correlations across countries, it is not surprising that the model fit does not change when constraining factor variance and covariances in Model 4.

In summary, it can be concluded that for self-concept constructs higher levels of parameter invariance are found than for the constructs related to science teaching styles. This is not surprising given the fact that science teaching varies considerably across participating countries and that responses to these questions are likely to be affected by differences in teaching practices and the structure of the science curriculum. The rules for deciding at what degree of parameter invariance constructs should be accepted in international comparative research remain unclear. Tests of significance tend to indicate that less constrained models always fit the population better than more constrained ones. Therefore, it is necessary to base judgements on relative model fits when comparing models with different levels of constraints. At the field trial stage (as is the case with data presented in this paper), analyses of cross-country may help to identify items and constructs that are more likely to exhibit measurement equivalence than others.

Different IRT Partial Credit models for self-concept items can be compared using the deviance statistic. The deviance is a statistic that indicates how well the item response model fits the data. When comparing the fit for two different models, this value can be compared to a Chi-Square distribution where the degrees of freedom are equal to the difference in the number of parameters to be estimated for each model. As with chi square tests in CFA, even smaller differences may appear as significant with larger sample size as usually used in international studies.

**Table 8 IRT Models with different constraints for self-concept items***

| | **SCSCIE** | | **SCMATH** | | **SCREAD** | |
|---|---|---|---|---|---|---|
| | **Deviance** | **Diff.** | **Deviance** | **Diff.** | **Deviance** | **Diff.** |
| Model1 | 159,037 (19) | | 150,049 (19) | | 137,788 (19) | |
| Model2 | 156,029 (61) | 3,007 | 148,172 (61) | 1,877 | 136,017 (61) | 1,771 |
| Model3 | 155,036 (109) | 994 | 147,405 (109) | 767 | 135,407 (109) | 610 |
| Model4 | 153,222 (289) | 1,814 | 145,637 (289) | 1,768 | 133,969 (289) | 1,437 |

* Number of estimated parameters in parentheses. Model 1 includes parameters for items and steps, Model 2 additional parameters for country group effect and group-item interaction, Model 3 additional parameters for country effect and country-item interaction, Model 4 additional parameters for country effect and country-item interaction and country-step interaction.

Table 8 shows deviance statistics, numbers of estimated parameters and the differences in deviances for each of the self-concept scales. Given the large sample size, it is not surprising that all the differences in deviances appear to be statistically significant. The largest difference in deviance for each of the constructs can be observed between Models 1 and 2. This means that introducing the effect of groups of countries on the item parameters makes a difference in model fit. Introducing a country effect instead of a country group (Model 3) effect makes further difference in overall model fit (it should be noted that the variation between groups of countries is included in the between-country variation). Estimating country-specific step parameters again leads to a reduction in deviance (Model 4).

**Table 9 IRT Models with different constraints for science teaching items**

| | HANDSON | | SCINVEST | | SCMODEL | |
|---|---|---|---|---|---|---|
| | Deviance | Diff. | Deviance | Diff. | Deviance | Diff. |
| Model1 | 134,250 (13) | | 119,665 (13) | | 160,946 (16) | |
| Model2 | 131,559 (41) | 2,692 | 117,003 (41) | 2,662 | 158,591 (51) | 2,354 |
| Model3 | 130,531 (73) | 1,028 | 116,341 (73) | 662 | 158,062 (91) | 529 |
| Model4 | 129,066 (193) | 1,464 | 115,681 (193) | 660 | 156,844 (241) | 1,219 |

* Number of estimated parameters in parentheses. Model 1 includes parameters for items and steps, Model 2 additional parameters for country group effect and group-item interaction, Model 3 additional parameters for country effect and country-item interaction, Model 4 additional parameters for country effect and country-item interaction and country-step interaction.

Table 9 shows the model comparison for the teaching style scales. As for self-concept scales, the largest reduction in deviance can be observed once country group effects are included in the model. As with the self-concept scales, all differences are statistically significant.

At the item level it is interesting to review to what extent individual item parameters vary across countries. This can be illustrated by looking at the effect parameters (CNT*ITEM) which indicate the extent to which different item parameters would have been estimated when calibrating for each country separately. In order to summarise these results for both sets of constructs, the average effects across all items within each scale are presented in Table 10 by country. Cells contain the average of the absolute values of the estimated effects across items. The estimated country effects are shown in detail in Appendix 2.

**Table 10 Average country effects* across items by country**

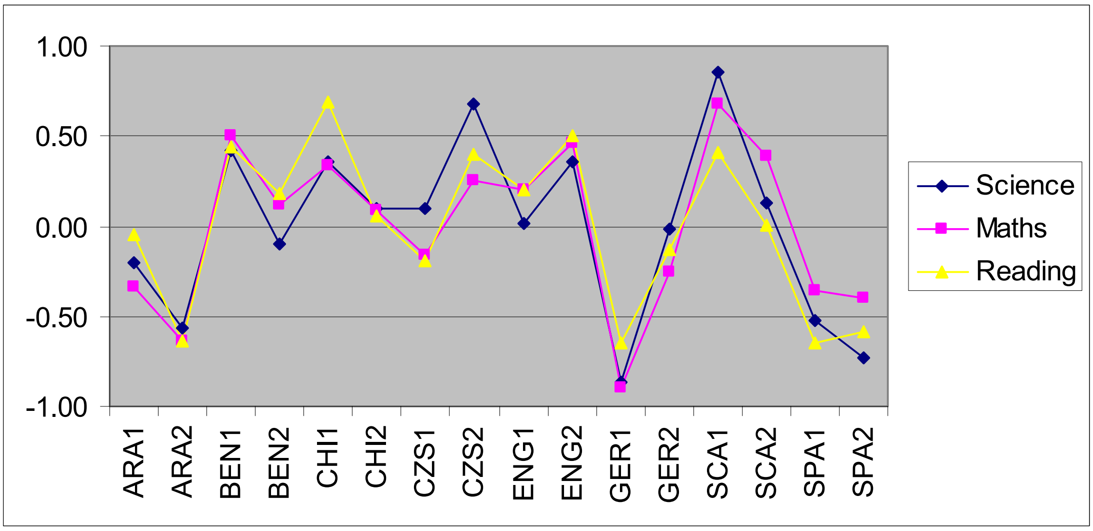
| Country | Self-concept scales | | | Teaching style scales | | |
|---|---|---|---|---|---|---|
| Country | SCSCIE | SCMATH | SCREAD | HANDSON | SCINVEST | SCMODEL |
| ARA1 | 0.42 | 0.59 | 0.34 | 0.10 | 0.02 | 0.15 |
| ARA2 | 0.20 | 0.22 | 0.22 | 0.33 | 0.26 | 0.16 |
| BEN1 | 0.26 | 0.23 | 0.19 | 0.26 | 0.10 | 0.08 |
| BEN2 | 0.18 | 0.22 | 0.21 | 0.29 | 0.30 | 0.18 |
| CHI1 | 0.35 | 0.43 | 0.50 | 0.27 | 0.22 | 0.27 |
| CHI2 | 0.34 | 0.39 | 0.40 | 0.29 | 0.15 | 0.11 |
| CZS1 | 0.20 | 0.24 | 0.33 | 0.32 | 0.25 | 0.21 |
| CZS2 | 0.23 | 0.22 | 0.30 | 0.35 | 0.25 | 0.28 |
| ENG1 | 0.13 | 0.17 | 0.12 | 0.11 | 0.22 | 0.18 |
| ENG2 | 0.14 | 0.16 | 0.19 | 0.18 | 0.22 | 0.15 |
| GER1 | 0.35 | 0.32 | 0.27 | 0.22 | 0.16 | 0.11 |
| GER2 | 0.13 | 0.20 | 0.22 | 0.33 | 0.11 | 0.19 |
| SCA1 | 0.38 | 0.25 | 0.22 | 0.23 | 0.42 | 0.18 |
| SCA2 | 0.27 | 0.29 | 0.33 | 0.13 | 0.15 | 0.12 |
| SPA1 | 0.38 | 0.22 | 0.27 | 0.30 | 0.17 | 0.15 |
| SPA2 | 0.33 | 0.39 | 0.36 | 0.20 | 0.19 | 0.07 |
| **Median** | **0.26** | **0.24** | **0.27** | **0.26** | **0.20** | **0.16** |

* The average country effect is the mean of absolute values of item-country interactions across items.

Most average country effects on item location parameters are between 0.15 and 0.30 logits, only the items measuring SCMODEL have (mostly) lower values. In some countries higher country effects can be observed. In particular for self-concept there are larger average effects (> 0.3) in the two Chinese-speaking countries, in one Arabic-speaking country and one Spanish-speaking country. The scale with least country effects on item location is SCMODEL (models and applications in science teaching). This is somewhat surprising given the very different contexts for teaching styles across countries.

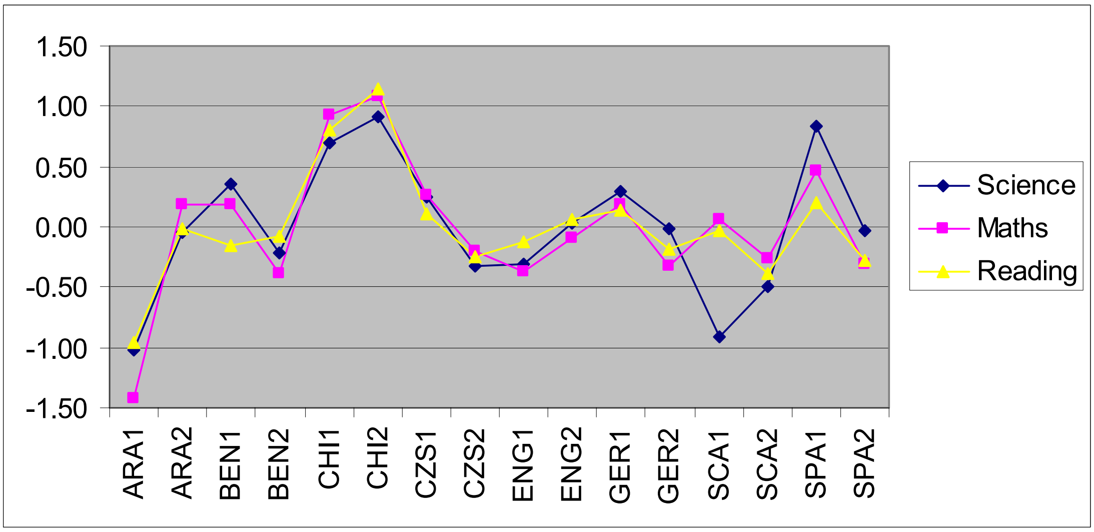
When looking at the average country effect of items (calculated as the mean of absolute values across countries per item) it can be observed that particular items show larger country effects than others (see Appendix 2). For the self-concept measures the location of the first two items in each scale (related to science, mathematics and reading) varies quite substantially across countries.

**Figure 1 Country effects on first self-concept item***



* IRT model: ITEM-CNT+ITEM*CNT+ITEM*STEP. Parameters are ITEM*CNT. Item wording is: "Learning advanced <science/mathematics/test language topics> would be easy for me".

**Figure 2 Country effects on second self-concept item***



* IRT model: ITEM-CNT+ITEM*CNT+ITEM*STEP. Parameters are ITEM*CNT. Item wording is: "I can usually give good answers to <test questions> on <science/mathematics/test language topics>".

Figure 1 and Figure 2 show the country effects on the location of these two items. It becomes obvious that regardless of the subject domain parameter variation follows similar patterns across countries. Clearly, the specific wording of these items and not the subject area as a reference point causes deviations from the item location for the pooled sample. Probably the reference to "advanced topics" (first item) and "test questions" (second items) has different meanings across the countries included in this analysis.

The IRT analyses show that there is a substantial variation in item parameters for the partial credit model across countries. From the analyses undertaken for this paper country effects on questionnaire item location parameters are found in the range of 0.1 to 0.3 logits. For some items, larger country effects are found, some of them seem to be associated with particular problems in wording (for example for the first two self-concept items).

However, it still needs to be determined at what point lack of parameter invariance becomes problematic. Rules of thumb can help to identify problems with parameter invariance and flag particular items or scales at the stage of field trials in order to inform the selection process for the main study. It needs to be recognised that some variation in model parameters is an inevitable characteristic of cross-national research and that researchers cannot expect their measures to be completely invariant given the numerous factors (linguistic and cultural differences, variation in teaching and learning across educational systems) that affect student responses to questionnaire items.

# Discussion

This paper shows how two different factor analytical methods can be used to detect parameter invariance in questionnaire items used in international surveys:

- Confirmatory Factor Analysis (CFA) is based on the analysis of covariance structures and allows researchers to test whether model structures hold within and across groups. Problems with non-normality in categorical items can be overcome by using poly-choric correlations (with WLSMV estimation) instead of covariances (with ML estimation).
- Item Response Theory (IRT) is based on the analysis of item responses and allows researchers to estimate group-specific effects on item model parameters.

It should be noted that one should not expect the same results when using both methods with the same data.[8] According to the analyses presented in this paper, CFA multiple-group modelling showed higher levels of parameter invariance for self-concept items than for science teaching items. The results for IRT models, on the other hand, indicate that country effects on item location parameters were larger for self-concept items than for the science teaching items.

In general, the analyses of PISA 2006 field trial data presented as examples in this paper show that there is evidence of parameter variance across languages, cultures and individual countries in international studies. Models that take group-specific variations in parameters into account generally fit the data better than constrained models where parameters are assumed to be equal. Constraining parameters within groups of countries typically leads to moderate increases in model fit but model fit

[8] Low item reliabilities (lack of variance explanation for items) in CFA, however, often correspond to item misfit when estimating IRT models.

usually improves further when taking the variation between individual countries into account.

Among the CFA or IRT models estimated in the analyses for this paper those with equal models parameters across countries were certainly the "least fitting". However, this is to be expected as additional parameters usually lead to improvements in model fit. Decisions about acceptance of cross-national construct validity and item-specific parameter variation should generally be based on "relative" judgement regarding model fit rather than stringent hypothesis testing. Using tests of statistical significance just shows that differences are not due to random noise but (given the larger samples sizes typically used in international research) do not provide proper guidance for what degree of parameter variation is within "acceptable limits".

It would be rather naive to assume that questions, translated from a source version into many different languages spoken in countries with very different cultures and educational systems, would be answered in exactly the same way. Some degree of parameter variation can be regarded as an inherent characteristic of cross-national research and the question is rather at what point this leads to measurement non-equivalence.  Therefore, the crucial question is at what point parameter variation becomes critical and leads to biased results.

When looking at the item dimensionality within countries separately, the examples have shown generally similar results across countries. Similar constructs are measured across countries but there is some parameter variation between countries in loadings (in CFA) and item locations (in IRT). It is unlikely that using these questionnaire measures would provide substantially different results when analysing at their relationships with other variables. However, the results of these analyses caution against simple comparisons of average scores across countries (for instance in lead tables).

Reviews of construct validity and measurement equivalence for questionnaire data have often been neglected in international educational research and similar to the analysis of test items they should be implemented in all international studies. When analysing questionnaire data in international studies, both CFA multiple-group models and IRT modelling of group effects on parameter estimates are helpful tools to scrutinise parameter invariance of constructs and items. This can be particularly useful at the field trial stage in order to select those items and constructs with higher levels of measurement equivalence. Looking at the extent to which parameter variation between countries is influenced by language and culture (by grouping countries according to common characteristics) may provide additional information in order to judge the causes of measurement non-equivalence for particular items and/or constructs.

# References

Adams, R. (2002). Scaling PISA cognitive data. In: Adams, R. and Wu, M. (eds.). *PISA 2000. Technical Report* (pp. 99-108). Paris: OECD Publications.

Andersen, Erling B. (1997). The Rating Scale Model. In: van der Linden, W. J. and Hambleton, R. K. (Eds.). *Handbook of Modern Item Response Theory* (pp. 67-84). New York/Berlin/Heidelberg: Springer.

Andrich, D. and Luo, G. (2003). Measuring Attitudes by Unfolding a Likert-style Questionnaire. In: J. Keeves & R. Watanabe (Eds.). *International Handbook of Educational Research in the Asia-Pacific Region* (pp. 1271-1283). Dordrecht: Kluwer Academic Publishers.

Bollen, K.A. and Long, S. J. (1993) (Eds.). *Testing Structural Equation Models*, Newbury Park/London.

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

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

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

Douglas, S. P. and Nijssen, E. J. (2003). On the use of "borrowed" scales in cross-national research: A cautionary note. *International Marketing Review*, 20:6, 621-642.

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

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

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

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

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

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

Maris, G. and Maris, E. (2002). *Are Attitude Items Monotone or Single-Peaked? An Analysis using Bayesian Methods.* (unpublished research paper). Arnheim: Citogroup. [http://download.citogroep.nl/pub/pok/reports/Report02-02.pdf]

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

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

Muthen, B. O., du Toit, S. H. C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Unpublished manuscript.

Organisation for Economic Cooperation and Development (2005). *Technical Report for the OECD Programme for International Student Assessment 2003*. Paris: OECD Publications.

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

Satorra, A. (1992). Asymptotic robust inference in the analysis of mean and covariance structures. In: P. V. Marsden (Ed.), *Sociological Methodology* (pp. 249-278). Oxford: Blackwell.

Schulz, W. (2002). Constructing and Validating Questionnaire Indices. In: Adams, R. and Wu, M. (Ed.). *Technical Report for the OECD Programme for International Student Assessment* (pp. 217-252). Paris: OECD Publications.

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

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

Schulz, W. and Sibberns, H. (2004). Scaling Procedures for Cognitive Items. In: W. Schulz and H. Sibberns (eds.) *IEA Civic Education Study. Technical Report* (pp. 69-91). Amsterdam: IEA.

Smith, T. W. (2004). Developing and Evaluating Cross-National Survey Instruments. In: S. Presser et. al. (eds.). *Methods for Testing and Evaluation Survey Questionnaires* (pp. 431-453). Hoboken: Wiley.

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

Warm, T. A. (1989). Weighted Likelihood Estimation of Ability in Item Response Theory. *Psychometrika*, 54(3), 427-450.

Wilson, M. (1994). Comparing Attitude Across Different Cultures: Two Quantitative Approaches to Construct Validity. In: M. Wilson (Ed.), *Objective measurement II: Theory into practice* (pp. 271-292). Norwood, NJ: Ablex.

Wu, M.L., Adams, R.J., and Wilson, M.R. (1997). *ConQuest: Multi-Aspect Test Software* [computer program manual]. Camberwell, Vic.: Australian Council for Educational Research.

# Appendix 1 Multiple-group model results

**Table 11 Results for multiple-group models (self-concept items)**

| TESTS OF MODEL FIT* | Constrained Covariances, variances and loadings | Constrained factor loadings across countries | Constrained factor loadings within groups of countries | Unconstrained Model |
|---|---|---|---|---|
| **Chi-Square Test of Model Fit** | | | | |
| **Value** | 2854 | 8040 | 7321 | 6836 |
| **Degrees of Freedom** | 285 | 1006 | 961 | 902 |
| **CFI** | 0.99 | 0.96 | 0.97 | 0.97 |
| **TLI** | 0.99 | 0.99 | 0.99 | 0.99 |
| **Number of Free Parameters** | 390 | 480 | 578 | 705 |
| **RMSEA** | 0.103 | 0.091 | 0.088 | 0.088 |
| **WRMR** | 10.776 | 7.356 | 7.021 | 6.789 |

* RMSEA = Root Mean Square Error Of Approximation; WRMR = Weighted Root Mean Square Residual.

**Table 12 Results for multiple-group models (science teaching items)**

| TESTS OF MODEL FIT* | Constrained Covariances, variances and loadings | Constrained factor loadings across countries | Constrained factor loadings within groups of countries | Unconstrained Model |
|---|---|---|---|---|
| **Chi-Square Test of Model Fit** | | | | |
| **Value** | 6968 | 9807 | 7856 | 7302 |
| **Degrees of Freedom** | 682 | 944 | 897 | 846 |
| **CFI** | 0.86 | 0.80 | 0.84 | 0.85 |
| **TLI** | 0.93 | 0.93 | 0.94 | 0.94 |
| **Number of Free Parameters** | 295 | 385 | 455 | 535 |
| **RMSEA** | 0.105 | 0.106 | 0.097 | 0.096 |
| **WRMR** | 10.565 | 9.641 | 8.552 | 8.134 |

* RMSEA = Root Mean Square Error Of Approximation; WRMR = Weighted Root Mean Square Residual.

# Appendix 2 Country effects on item parameters

## Table 13 Estimated IRT country effects on item parameters (SCSCIE)*

| Country | ST20Q01 | ST20Q02 | ST20Q03 | ST20Q04 | ST20Q05 | ST20Q06 |
|---|---|---|---|---|---|---|
| ARA1 | -0.21 | -1.02 | 0.38 | 0.23 | -0.05 | 0.66 |
| ARA2 | -0.57 | -0.04 | 0.31 | 0.25 | 0.00 | 0.05 |
| BEN1 | 0.42 | 0.35 | -0.01 | -0.08 | -0.49 | -0.18 |
| BEN2 | -0.10 | -0.21 | -0.12 | 0.33 | 0.22 | -0.11 |
| CHI1 | 0.36 | 0.69 | -0.19 | -0.24 | -0.32 | -0.31 |
| CHI2 | 0.10 | 0.91 | -0.15 | -0.08 | -0.18 | -0.60 |
| CZS1 | 0.10 | 0.25 | 0.11 | 0.09 | -0.60 | 0.06 |
| CZS2 | 0.68 | -0.33 | -0.16 | -0.03 | -0.06 | -0.11 |
| ENG1 | 0.02 | -0.31 | 0.13 | 0.11 | 0.13 | -0.08 |
| ENG2 | 0.36 | 0.04 | 0.00 | -0.17 | 0.02 | -0.24 |
| GER1 | -0.87 | 0.30 | -0.11 | 0.15 | 0.61 | -0.08 |
| GER2 | -0.02 | -0.02 | -0.15 | -0.21 | 0.05 | 0.34 |
| SCA1 | 0.86 | -0.91 | -0.09 | 0.01 | -0.16 | 0.28 |
| SCA2 | 0.13 | -0.50 | -0.04 | -0.27 | 0.36 | 0.32 |
| SPA1 | -0.53 | 0.84 | 0.10 | 0.13 | 0.08 | -0.62 |
| SPA2 | -0.73 | -0.03 | 0.00 | -0.23 | 0.38 | 0.61 |
| **Average** | **0.38** | **0.42** | **0.13** | **0.16** | **0.23** | **0.29** |
| ***Item parameter*** | ***0.50*** | ***-0.50*** | ***-0.07*** | ***0.55*** | ***-0.37*** | ***-0.12*** |

* ConQuest IRT model: ITEM-CNT+ITEM*CNT+ITEM*STEP. Parameters are ITEM*CNT. The average country effect per item is calculated as the mean of absolute values across countries.

## Table 14 Estimated IRT country effects on item parameters (SCMATH)*

| Country | ST21Q01 | ST21Q02 | ST21Q03 | ST21Q04 | ST21Q05 | ST21Q06 |
|---|---|---|---|---|---|---|
| ARA1 | -0.34 | -1.42 | 0.73 | 0.26 | 0.33 | 0.44 |
| ARA2 | -0.64 | 0.19 | 0.26 | 0.19 | -0.02 | 0.02 |
| BEN1 | 0.50 | 0.18 | -0.11 | -0.07 | -0.53 | 0.03 |
| BEN2 | 0.12 | -0.38 | -0.05 | 0.38 | -0.21 | 0.15 |
| CHI1 | 0.34 | 0.94 | -0.28 | -0.23 | -0.37 | -0.40 |
| CHI2 | 0.08 | 1.09 | -0.29 | -0.21 | -0.28 | -0.40 |
| CZS1 | -0.16 | 0.27 | 0.09 | 0.23 | -0.55 | 0.13 |
| CZS2 | 0.25 | -0.20 | -0.30 | 0.11 | 0.29 | -0.15 |
| ENG1 | 0.21 | -0.37 | 0.09 | -0.06 | 0.21 | -0.08 |
| ENG2 | 0.47 | -0.09 | -0.03 | -0.17 | -0.01 | -0.17 |
| GER1 | -0.90 | 0.19 | -0.04 | 0.44 | 0.32 | -0.02 |
| GER2 | -0.26 | -0.33 | -0.03 | 0.08 | 0.07 | 0.47 |
| SCA1 | 0.68 | 0.06 | 0.01 | -0.12 | -0.14 | -0.49 |
| SCA2 | 0.39 | -0.27 | -0.28 | -0.31 | 0.31 | 0.16 |
| SPA1 | -0.35 | 0.46 | -0.08 | -0.05 | 0.22 | -0.19 |
| SPA2 | -0.40 | -0.31 | 0.31 | -0.46 | 0.34 | 0.51 |
| **Average** | **0.38** | **0.42** | **0.19** | **0.21** | **0.26** | **0.24** |
| ***Item parameter*** | ***0.78*** | ***-0.47*** | ***-0.23*** | ***0.51*** | ***-0.53*** | ***-0.05*** |

* ConQuest IRT model: ITEM-CNT+ITEM*CNT+ITEM*STEP. Parameters are ITEM*CNT. The average country effect per item is calculated as the mean of absolute values across countries.

**Table 15 Estimated IRT country effects on item parameters (SCREAD)***

| Country | ST22Q01 | ST22Q02 | ST22Q03 | ST22Q04 | ST22Q05 | ST22Q06 |
|---|---|---|---|---|---|---|
| ARA1 | -0.05 | -0.95 | 0.26 | -0.02 | 0.25 | 0.51 |
| ARA2 | -0.64 | -0.02 | 0.21 | 0.22 | -0.02 | 0.25 |
| BEN1 | 0.44 | -0.16 | -0.25 | -0.06 | -0.10 | 0.12 |
| BEN2 | 0.18 | -0.08 | -0.34 | -0.05 | -0.15 | 0.44 |
| CHI1 | 0.69 | 0.81 | -0.28 | -0.32 | -0.45 | -0.45 |
| CHI2 | 0.06 | 1.14 | -0.40 | -0.02 | -0.37 | -0.41 |
| CZS1 | -0.19 | 0.11 | 0.39 | 0.49 | -0.51 | -0.30 |
| CZS2 | 0.40 | -0.24 | -0.12 | 0.50 | -0.11 | -0.42 |
| ENG1 | 0.20 | -0.12 | 0.03 | 0.14 | -0.04 | -0.21 |
| ENG2 | 0.51 | 0.07 | -0.15 | -0.01 | -0.09 | -0.33 |
| GER1 | -0.65 | 0.14 | -0.08 | 0.26 | 0.41 | -0.08 |
| GER2 | -0.13 | -0.19 | -0.16 | -0.18 | 0.25 | 0.41 |
| SCA1 | 0.41 | -0.03 | 0.24 | -0.01 | -0.35 | -0.26 |
| SCA2 | 0.01 | -0.39 | 0.43 | -0.61 | 0.37 | 0.19 |
| SPA1 | -0.64 | 0.20 | -0.01 | -0.14 | 0.51 | 0.09 |
| SPA2 | -0.59 | -0.27 | 0.24 | -0.21 | 0.39 | 0.45 |
| **Average** | **0.36** | **0.31** | **0.22** | **0.20** | **0.27** | **0.31** |
| ***Item parameter*** | ***0.44*** | ***-0.24*** | ***-0.16*** | ***0.42*** | ***-0.41*** | ***-0.06*** |

* ConQuest IRT model: ITEM-CNT+ITEM*CNT+ITEM*STEP. Parameters are ITEM*CNT. The average country effect per item is calculated as the mean of absolute values across countries.

**Table 16 Estimated IRT country effects on item parameters (HANDSON)***

| Country | ST36Q01 | ST36Q09 | ST36Q11 | ST36Q14 |
|---|---|---|---|---|
| ARA1 | -0.10 | 0.09 | -0.10 | 0.11 |
| ARA2 | -0.10 | 0.42 | -0.56 | 0.24 |
| BEN1 | -0.52 | 0.01 | 0.08 | 0.43 |
| BEN2 | 0.33 | -0.12 | -0.46 | 0.25 |
| CHI1 | -0.23 | 0.05 | 0.49 | -0.31 |
| CHI2 | -0.37 | 0.53 | -0.20 | 0.04 |
| CZS1 | 0.45 | 0.14 | 0.06 | -0.65 |
| CZS2 | 0.47 | -0.19 | 0.23 | -0.51 |
| ENG1 | -0.12 | -0.10 | 0.20 | 0.02 |
| ENG2 | -0.09 | -0.26 | 0.32 | 0.03 |
| GER1 | 0.03 | -0.01 | -0.43 | 0.41 |
| GER2 | 0.27 | 0.02 | -0.66 | 0.37 |
| SCA1 | -0.43 | 0.09 | 0.37 | -0.04 |
| SCA2 | -0.12 | -0.15 | 0.19 | 0.08 |
| SPA1 | 0.41 | -0.18 | 0.19 | -0.42 |
| SPA2 | 0.12 | -0.35 | 0.28 | -0.05 |
| **Average** | **0.26** | **0.17** | **0.30** | **0.25** |
| ***Item parameter*** | ***0.91*** | ***-0.41*** | ***0.08*** | ***-0.58*** |

* ConQuest IRT model: ITEM-CNT+ITEM*CNT+ITEM*STEP. Parameters are ITEM*CNT. The average country effect per item is calculated as the mean of absolute values across countries.

### Table 17 Estimated IRT country effects on item parameters (SCINVEST)*

| **Country** | **ST36Q03** | **ST36Q05** | **ST36Q12** | **ST36Q16** |
|---|---|---|---|---|
| ARA1 | -0.03 | -0.01 | 0.01 | 0.04 |
| ARA2 | 0.23 | -0.18 | -0.33 | 0.28 |
| BEN1 | -0.13 | 0.20 | -0.06 | -0.01 |
| BEN2 | 0.59 | -0.31 | -0.29 | 0.01 |
| CHI1 | 0.32 | 0.13 | -0.39 | -0.06 |
| CHI2 | 0.01 | 0.26 | 0.03 | -0.30 |
| CZS1 | 0.07 | -0.50 | 0.32 | 0.10 |
| CZS2 | 0.14 | 0.09 | 0.27 | -0.50 |
| ENG1 | -0.43 | 0.01 | 0.24 | 0.18 |
| ENG2 | -0.45 | 0.21 | 0.14 | 0.10 |
| GER1 | -0.14 | 0.21 | -0.18 | 0.11 |
| GER2 | 0.22 | -0.11 | -0.11 | 0.00 |
| SCA1 | -0.77 | -0.06 | 0.38 | 0.46 |
| SCA2 | -0.03 | -0.14 | -0.12 | 0.29 |
| SPA1 | 0.14 | 0.17 | 0.03 | -0.34 |
| SPA2 | 0.28 | 0.04 | 0.06 | -0.38 |
| **Average** | **0.25** | **0.16** | **0.19** | **0.20** |
| ***Item parameter*** | ***0.01*** | ***0.40*** | ***-0.10*** | ***-0.30*** |

* ConQuest IRT model: ITEM-CNT+ITEM*CNT+ITEM*STEP. Parameters are ITEM*CNT. The average country effect per item is calculated as the mean of absolute values across countries.

### Table 18 Estimated IRT country effects on item parameters (SCMODEL)*

| **Country** | ST36Q04 | ST36Q07 | ST36Q13 | ST36Q15 | ST36Q17 |
|---|---|---|---|---|---|
| ARA1 | -0.11 | -0.12 | 0.24 | -0.15 | 0.14 |
| ARA2 | 0.03 | 0.36 | 0.01 | -0.30 | -0.10 |
| BEN1 | -0.12 | -0.05 | 0.18 | -0.04 | 0.03 |
| BEN2 | -0.43 | 0.00 | -0.02 | 0.39 | 0.05 |
| CHI1 | -0.36 | -0.33 | 0.45 | 0.10 | 0.14 |
| CHI2 | -0.09 | -0.18 | 0.20 | 0.08 | -0.02 |
| CZS1 | -0.06 | 0.02 | 0.48 | -0.46 | 0.02 |
| CZS2 | 0.59 | -0.32 | 0.12 | -0.14 | -0.25 |
| ENG1 | 0.31 | 0.01 | -0.44 | 0.13 | -0.02 |
| ENG2 | 0.31 | 0.07 | -0.32 | 0.01 | -0.07 |
| GER1 | -0.05 | -0.04 | -0.20 | 0.23 | 0.06 |
| GER2 | -0.28 | 0.15 | -0.19 | 0.20 | 0.13 |
| SCA1 | 0.32 | 0.05 | -0.41 | -0.05 | 0.10 |
| SCA2 | 0.12 | -0.02 | -0.17 | 0.18 | -0.12 |
| SPA1 | -0.28 | 0.31 | 0.08 | -0.10 | 0.00 |
| SPA2 | 0.09 | 0.08 | -0.03 | -0.06 | -0.08 |
| **Average** | **0.22** | **0.13** | **0.22** | **0.16** | **0.08** |
| ***Item parameter*** | ***-0.18*** | ***0.57*** | ***0.05*** | ***-0.52*** | ***0.07*** |

* ConQuest IRT model: ITEM-CNT+ITEM*CNT+ITEM*STEP. Parameters are ITEM*CNT. The average country effect per item is calculated as the mean of absolute values across countries.