Validating Questionnaire Constructs in International Studies. Two Examples from PISA 2000

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 Chicago, 21-25 April 2003.

ABSTRACT

One of the most salient requirements for international educational research is the use of comparable measures. For the comparison of student performance across countries the use of IRT scaling techniques facilitates the collection of cross-nationally comparable measures. But there is also a need for valid and comparable context variables, such as family background, learning context, motivational factors and so on. A wide range of student and school information was gathered through the PISA student and school questionnaires. Most of the theoretical constructs were measured through sets of items that needed to be validated across countries. After the process of construct validation, IRT modelling was used to scale the items. This method not only provides a more sophisticated scaling technique but also an elegant way of dealing with incomplete data (missing values). This paper describes the process of cross-country validation in the PISA study with two examples and discusses its limitations and problems.

INTRODUCTION

One of the important challenges of international educational research is the search for comparable measures of student background, attitudes and perceptions. This paper describes methodological approaches for validating constructs derived from the student and school questionnaires used in the PISA study, and discusses their limitations and problems. Cross-country validity of these constructs is of particular importance as measures derived from questionnaires are often used to explain differences in student performance within and across countries and are, thus, potential sources of policy-relevant information about ways of improving educational systems.

The use of instruments across national and cultural groups requires not only a thorough process of translation into different languages and its verification, it also makes assumptions about having measured similar characteristics, attitudes and perceptions in different national and cultural contexts.

Wilson (1994) suggests that in addition to the need for appropriate translation, psychometric techniques should be used to analyse the extent to which constructs have (1) *consistent dimensionality* and (2) *consistent construct validity* in different national and cultural contexts, that is, once the measurement stability for each subscale is confirmed, also the multidimensional relationship between these subscales should be confirmed. This was illustrated based on student data from Australia and the United States by assessing the construct validity of three scales derived from Likert-type items using Structural Equation Modelling (SEM) and Item Response Theory (IRT). It could be shown that whereas with the SEM approach the validity of these measures could be largely confirmed, the IRT approach showed more discrepancies between the two samples.

Another approach to review construct validity is to analyse the relationship of a measure with other (related) 'reference' variables. This is limited by two important constraints:

- These 'reference' variables might have similar problems of validity and/or reliability.
- Often the relationship of the measure with a 'reference' variable is an empirical question in itself, that is, researchers must be aware that a low or negative correlation could be explained by factors other than missing construct validity.

This paper will demonstrate different approaches to the cross-country validation of Questionnaire constructs derived from the PISA questionnaires and will discuss problems and limitations as well as future perspectives.

QUESTIONNAIRE CONSTRUCTS IN PISA

In each cycle of the PISA study a questionnaire is used to collect information on student background, interest and engagement, instructional practice, and indicators of school and classroom climate. In PISA 2000 a number of composites were derived from these questions that were used to report student characteristics and their relationship with student performance (see OECD 2001; Schulz, 2002). In addition to attempts to safeguard high standards for translation and cultural adaptation of the questionnaires through double-translation and independent verification (Grisay, 2002), an extensive analysis of item dimensionality was undertaken to validate the constructs across countries.

Conceptually, two types of PISA constructs can be distinguished:

- Constructs derived from Likert-type items used to measure perceptions, beliefs or attitudes.
- Constructs derived from factual statements about household possessions.

Whereas in the first type of constructs stimuli were developed by the researchers and provided to students for the purpose of measuring unobserved traits, for the second type of constructs factual statements were collected to derive indices about home background. Though for both types of constructs item dimensionality and internal scale consistency need to be assessed, there are differences to the extent these requirements can be met.

In the first category of constructs, sets of items are chosen that are uni-dimensional and have high internal consistency, discarding those items with unsatisfactory scaling properties. Here, both the stability of item dimensionality and item parameter invariance across countries are obvious criteria for validating these constructs.

For the second category of constructs, item development is more constrained by the factual nature of statements used for measurement. Additionally, the meaning of indicators is even more likely to be affected by cultural differences: Certain home possessions, for example, may be common in industrialised countries, but not in less developed countries. Furthermore, any kind of home possessions is likely to be affected by means of family income, but will also depend on other important variables (interest, cultural importance, habits) that influence its acquisition. Thus, assessment of item dimensionality needs to take this into account and researcher might consider less rigid standards for the uni-dimensionality of constructs. Here, a validation approach might give more weight to the analysis of correlates with related background variables and the predictive power.

This paper includes two examples of how construct validity in the PISA study can be assessed. One example is a set of Likert-type items about the perception of classroom climate measuring three different constructs, the other example a measure of family wealth derived from student reports about home possessions.

METHODOLOGICAL APPROACHES

Construct Validation based on Covariance Analysis

Structural Equation modelling (SEM) can be used to confirm theoretically expected dimensions and, if necessary, to re-specify the dimensional structure (Kaplan, 2000). The latter needs to be done carefully and should always be consistent with theory. Structural equation modelling takes the measurement error associated with the indicators into account and provides a tool for analysing the dimensional item structure and the estimation of the disattenuated correlation between latent variables.

For a Confirmatory Factor Analyses (CFA) with SEM an expected covariance matrix is fitted according to the theoretical factor structure. This can be done due to the possibility of computing the covariances from the estimates in the model and the estimated variance of the latent variables. Maximum Likelihood Estimation provides model estimates trying to minimise the differences between the expected (Σ) and the observed covariance matrix (S).[1]

Measures for the overall fit of a model then are obtained by comparing the expected Σ 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" and some re-specification may be necessary or, if this is not possible, the theoretical model has to be rejected. Assessment of model fit for SEM can be based on the following measures:

- The *Root Mean Square Error of Approximation* (RMSEA) measures the "discrepancy per degree of freedom for the model" (Browne and Cudeck, 1993: 144). A value of .05 and less is an indication of a close fit, values of .08 and more indicate a reasonable error of approximation and values greater than 1.0 typically lead to the rejection of a model.
- The *Goodness-of-Fit-Index* (GFI) is 'classical' measure of model fit. It measures the amount of variance in S explained with Σ and should be close to 1.0 to indicate a good model fit. As its distributional properties are unknown there are no standards for this fit measure.
- The *Root Mean Square Residual* (RMR) is a measure of the discrepancy between S and Σ, and is based on the residual matrix. Its values should be lower than .05 to indicate a good model fit.

---

[1] For ordinal variables Jöreskog and Sörbom (1993) recommend to use *Weighted Least Square Estimation* (WLS) with polychoric correlation matrices and corresponding asymptotic covariance weight matrices. Maximum Likelihood and Generalised Least Square (GLS) estimation both require normal distribution and continuous variables. However, as the main purpose of this kind of analyses is to analyse the dimensional structure of items, and Maximum Likelihood estimation provides robust estimates with respect to non-normality, its is deemed to be appropriate for the purpose described in this paper.

Item reliability, that is the amount of variance in an item explained by the latent variable, is another indicator of model fit. If the latent variables hardly explains any variance in one or more of the manifest variables the assumed factor structure is hardly confirmed even if the overall model fit still appears to be reasonable. This criterion can also be regarded as an indication of item fit where low item reliability typically leads to its deletion from a scale. Review of item reliabilities across countries is an indication of cross-country validity: If item reliability varies across countries, one cannot assume to have internationally comparable measures.

The *correlation between latent variables* gives an indication about the degree of similarity between the measured constructs, for example, whether a one- or two-dimensional model is more appropriate for a set of items. However, differences in correlations between latent variables across countries might be due to other causes than a lack of construct validity.

Scale reliability, typically measured using Cronbach's Alpha, gives an overall indication of how much variance in a scale can be contributed to the true score and how much of the variance is due to measurement error.

With SEM it is possible to estimate so-called 'Multiple Group Models' (see Kaplan, 2002, Chapter 4) where researcher may constrain parameters for a number of groups. A multiple group model is estimated based on the covariance matrices of all groups. By estimating common parameters it can be tested to what extent the same measurement model holds for all sub-groups. Fit indices like RMR and GFI can be compared between unconstrained and constrained models for each sub-group. In the case of international studies, country samples can be treated as sub-groups in such a multiple group model.

Construct Validation based on Item Response Theory (IRT)

Item Response Theory (IRT) was used for the scaling of the PISA 2000 questionnaire constructs. *Weighted Likelihood Estimates* (Warm, 1989) were computed and transformed into an international metric with an OECD mean of 0 and a standard deviation of 1.

In the case of categorical items with k categories the One-Parameter (Rasch) Model can be generalised to

$$P_{x_i}(q) = \frac{\exp(\sum_{j=0}^{x} q_n - d_i + t)}{1 + \exp(\sum_{j=1}^{k} q_n - d_i + t_{ij})}, \; x = 0, 1, 2, \ldots, m_i$$

where $P_{xi}(\theta)$ denotes the probability of person *n* to score *x* on item *i*. Here, τ_{ij} denotes an additional step parameter. For attitudinal items the parameter θ_n denotes the location of a person on the latent dimension. The item parameter δ_i gives the location of the item on the latent continuum, in the case of attitudinal items low values denote that an item is relatively easy to agree with, high values that an item is relatively hard

to agree with. The so-called Partial Credit Model (Masters and Wright, 1997) estimates step parameters for items of the same scale separately whereas the Rating Scale Model has the same step parameters for all items in a scale (Andersen, 1997).

IRT model fit can be assessed using Mean Square statistics (see Wright and Masters, 1982). The value of the item fit statistics should be close to 1.0 to indicate a good fit according to the model. Values greater than 1.0 indicate that the item discrimination is less, values less than 1.0 that the item discrimination is higher than expected. As the unweighted mean square residual (Outfit) statistic may be affected by a small number of outlying observations the weighted mean square residual (Infit) statistic is typically used as a criterion for assessing model fit.

To analyse scaling properties and to test parameter invariance in international studies the following approaches could be used:

- Review of item fit across and within countries to select items with satisfactory scaling properties.
- Comparison of item parameters across countries, to determine whether there is an item-by-country interaction, item parameters should have the same relative location in each country. This can be done both for the item location parameter δ_i and for step parameters τ_{ij}. A requirement of invariance for the step parameters would be a very strict standard for construct validity.
- Review of item fit after constraining item and step parameters to values derived from an international calibration sample. This is another test of parameter invariance, as item-by-country interaction leads automatically to item misfit for the respective country.

If the IRT model fits the data, person parameters for the latent dimension can be computed and used for subsequent analysis. Though highly correlated with the original raw scores, this method provides a sophisticated scaling method for dealing with missing values because estimates for the latent dimension may be obtained for all respondents who have answered at least one of the items.

The advantages of using IRT modelling for the scaling of PISA questionnaire data were the following: (1) It provides a tool for assessing scaling properties, (2) it allows to anchor item and step parameters in order to achieve comparable scores even for different samples and situations, where different items and a common set of core items is used, and (3) provides an elegant way of dealing with missing responses.

In PISA 2000 (also due to timeline restrictions) only limited use was made of the cross-national comparison of IRT scaling properties. Constructs were validated across countries mainly through Structural Equation Modelling and an assessment of scale reliabilities across countries. However, in future assessments more extensive use of the features of IRT scaling could be made to apply a more rigid test of parameter invariance across countries.

EXAMPLE 1: CLASSROOM CLIMATE

In PISA 2000 a set of 17 items was used to measure the classroom climate in test language lessons. 16 of these Likert-type items were used to derive three constructs: Disciplinary Climate (DISCI) describing problems with student discipline in the classroom (6 items),[2] Teacher Support (TSUP) describing the extent to which teachers are perceived by students as supportive (6 items), and Achievement Press (ACHPR) describing to what extent teachers are perceived as demanding by the students (4 items).

Table 1 shows the item numbering in the PISA 2000 student questionnaire, the item wording and their allocation to constructs.

Table 1: Items used to measure Classroom Climate

Q26	**How often do these things happen in your <test language> lessons?** *(Never or hardly ever, Some lessons, Most lessons, Every lesson)*
	Disciplinary Climate (DISCI)
e)	The teacher has to wait a long time for students to <quieten down>.
f)	Students cannot work well.
g)	Students don't listen to what the teacher says.
h)	Students don't start working for a long time after the lesson begins.
i)	There is noise and disorder.
j)	At the start of class, more than five minutes are spent doing nothing
	Teacher Support (TSUP)
a)	The teacher shows an interest in every student's learning.
l)	The teacher gives students an opportunity to express opinions.
m)	The teacher helps students with their work.
n)	The teacher continues teaching until the students understand.
p)	The teacher does a lot to help students.
q)	The teacher helps students with their learning.
	Achievement Press (ACHPR)
c)	The teacher wants students to work hard.
d)	The teacher tells students that they can do better.
e)	The teacher objects when students deliver <careless> work.
o)	Students have to learn a lot.

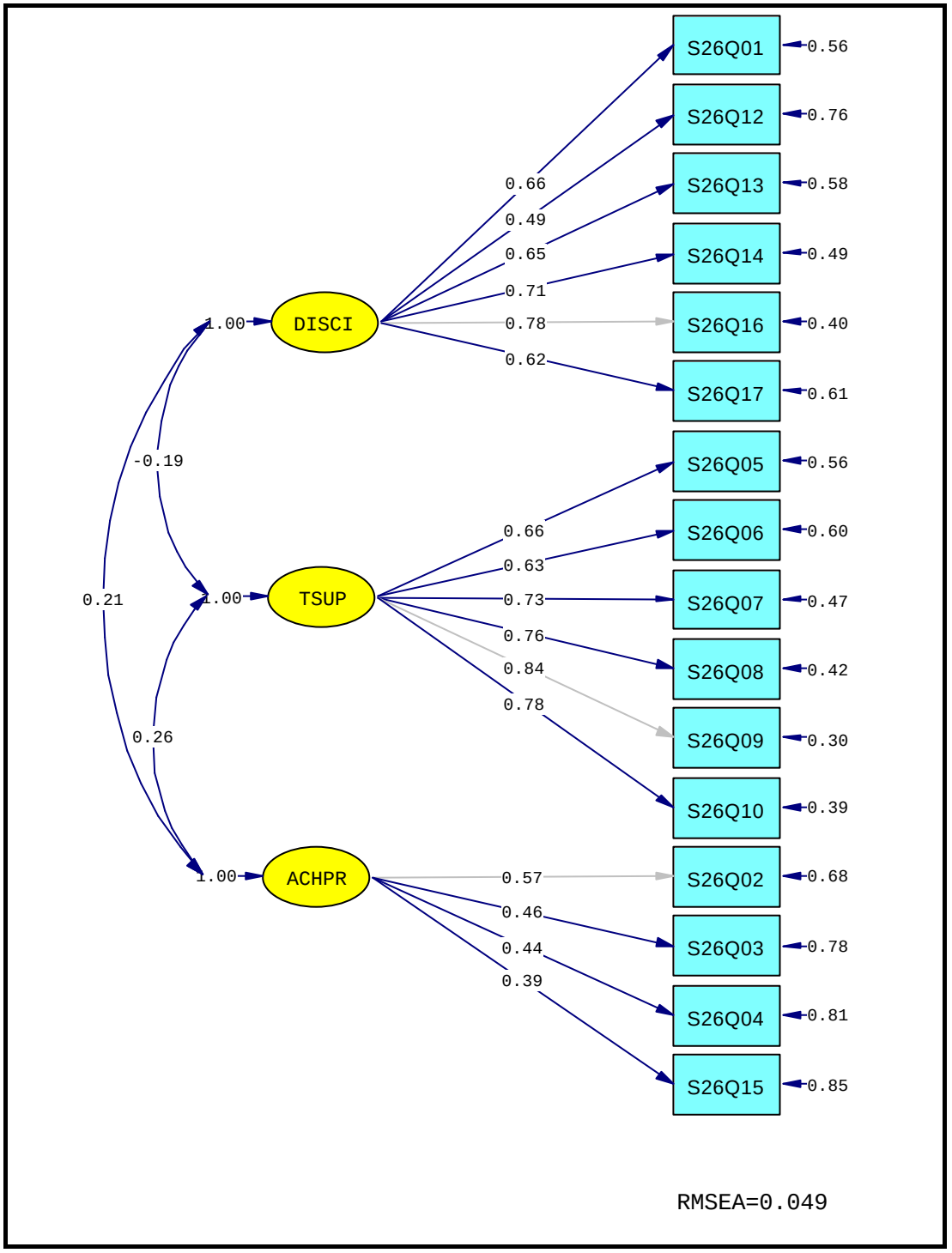
Based on an international calibration sample with 500 students randomly selected from each participating OECD country data set, in a first step a three-factor model was estimated using Structural Equation Modelling (LISREL). Graph 1 shows the standardised estimates for this model. Whereas DISCI and TSUP are negatively correlated (-.19), ACHPR has moderate positive correlations with both other latent variables (.21 with DISCI, .26 with TSUP).

The RMSEA fit index is satisfactory with .049 and most items measuring DISCI and TSUP have relatively strong loadings. However, the item loadings for ACHPR are generally lower and indicate a weaker measurement model.

---

2 Please note that in the first International Report (OECD, 2001), the scores were inverted so that low values indicated a poor disciplinary climate.

Graph 1: Three-Factor-Model for Classroom Climate Items



Standardised Estimates. Based on OECD calibration sample.

In a second step, for each country sub-sample the same model was estimated separately. In order to have comparable sample sizes and to avoid weighting, the sub-samples of the calibration sample each with 500 students were used for this purpose. A third step consisted of estimating a constrained model were item loadings were the same for each country sub-sample. Table 2 shows the fit indices for each country, RMR and GFI were used to compare the relative fit for the unconstrained versus the

constrained model.[3]

Table 2: Model fit across OECD PISA countries

	RMSEA	RMR		GFI	
	Model 1	Model 1	Model 2	Model 1	Model 2
Australia	0.062	0.045	0.049	0.93	0.92
Austria	0.052	0.057	0.064	0.94	0.93
Belgium	0.070	0.045	0.049	0.92	0.91
Canada	0.055	0.040	0.044	0.94	0.93
Czech Republic	0.065	0.042	0.046	0.92	0.92
Denmark	0.055	0.038	0.057	0.94	0.92
Finland	0.052	0.031	0.034	0.94	0.94
France	0.047	0.041	0.046	0.95	0.94
Germany	0.068	0.058	0.061	0.92	0.91
Greece	0.050	0.054	0.077	0.94	0.92
Hungary	0.066	0.055	0.063	0.92	0.90
Iceland	0.063	0.037	0.042	0.93	0.92
Ireland	0.056	0.052	0.060	0.94	0.93
Italy	0.063	0.045	0.056	0.93	0.92
Japan	0.073	0.058	0.094	0.91	0.88
Korea	0.063	0.054	0.100	0.93	0.87
Mexico	0.061	0.047	0.058	0.93	0.91
Netherlands	0.057	0.039	0.046	0.94	0.93
New Zealand	0.056	0.040	0.046	0.94	0.93
Norway	0.058	0.044	0.047	0.93	0.93
Poland	0.066	0.054	0.055	0.92	0.91
Portugal	0.056	0.031	0.041	0.94	0.92
Spain	0.056	0.042	0.046	0.94	0.93
Sweden	0.052	0.036	0.040	0.94	0.94
Switzerland	0.057	0.051	0.060	0.93	0.92
United Kingdom	0.051	0.041	0.046	0.94	0.94
United States	0.064	0.048	0.054	0.93	0.91

Model 1: unconstrained, Model 2: constrained item loadings. Based on OECD calibration sub-samples.

Generally, the fit of the separate models is satisfactory across countries, in no country the RMSEA indicates a poor item fit of >.08. However, constraining item loadings leads to weaker model fit in a number of countries, especially in Japan, Korea, and, to a lesser degree, in Greece.

[3] The RMSEA for the constrained model (2) is only available for the pooled data set.

Table 3: Scale Reliabilities for Classroom Climate Scales

	Reliabilities		
	DISCI	**TSUP**	**ACHPR**
Australia	0.84	0.89	0.52
Austria	0.82	0.87	0.54
Belgium	0.81	0.85	0.52
Canada	0.84	0.90	0.55
Czech Republic	0.79	0.78	0.67
Denmark	0.79	0.86	0.38
Finland	0.85	0.88	0.64
France	0.79	0.88	0.66
Germany	0.81	0.87	0.52
Greece	0.69	0.88	0.50
Hungary	0.82	0.87	0.53
Iceland	0.81	0.87	0.60
Ireland	0.85	0.90	0.49
Italy	0.82	0.85	0.51
Japan	0.83	0.91	0.40
Korea	0.78	0.81	0.24
Mexico	0.74	0.83	0.49
Netherlands	0.85	0.85	0.48
New Zealand	0.81	0.87	0.60
Norway	0.84	0.88	0.51
Poland	0.81	0.87	0.57
Portugal	0.74	0.87	0.49
Spain	0.81	0.88	0.59
Sweden	0.79	0.89	0.59
Switzerland	0.78	0.86	0.62
United Kingdom	0.85	0.89	0.40
United States	0.83	0.90	0.56
OECD average	**0.78**	**0.84**	**0.51**

Based on OECD calibration sub-samples. Reliabilities of less than two standard deviations below the OECD average marked yellow.

Table 3 shows the scale reliabilities (Cronbach's Alpha) within each sub-sample: TSUP has highly satisfactory reliabilities in all country sub-samples, the internal consistency for DISCI is notably lower only for Greece. ACHPR has generally rather poor scale reliabilities; especially Korea shows a very low scale reliability of .24 for this scale.

Problems with model fit can be reviewed in some detail by looking at the item reliabilities:

- For DISCI (Table A in the Appendix), it is interesting to note that the items h (‚Students don't start working for a long time after the lesson begins.') and j (‚At the start of class, more than five minutes are spent doing nothing.') have

very low reliabilities for Greece with 16 and 13 percent of explained item variance. According to information obtained from this country the item wording is problematic in the context of a particular educational system where lessons typically begin with a period without instruction.

- For ACHPR (see Table C in Appendix) in Japan and Korea three out of four items have low reliabilities. This means that the variance of the latent variable is mainly derived from one out of four items and explains the weakness of fit for the model with constrained item loadings. Also in other countries item reliabilities were extremely low. Furthermore, none of the four items has consistently high item reliabilities across countries.

Table 4: Correlations between Latent Classroom Climate Measures (LISREL estimates)

	Latent correlations between...		
	TSUP&DISCI	ACHPR&DISCI	ACHPR&TSUP
Australia	-.30	.19	.20
Austria	-.17	.32	.02
Belgium	-.20	.32	-.14
Canada	-.32	.27	.26
Czech Republic	.00	.40	-.13
Denmark	-.36	.04	.21
Finland	-.20	.26	.12
France	-.21	.31	.03
Germany	-.17	.24	.20
Greece	-.14	.17	.47
Hungary	-.16	.16	-.06
Iceland	-.25	.18	.25
Ireland	-.50	.28	.13
Italy	-.31	.18	.16
Japan	-.19	.02	.47
Korea	-.24	.06	.72
Mexico	-.16	.14	.49
Netherlands	-.22	.42	.14
New Zealand	-.24	.38	.10
Norway	-.30	.13	.32
Poland	-.09	.19	.14
Portugal	-.15	.21	.44
Spain	-.20	.23	.20
Sweden	-.21	.19	.11
Switzerland	-.13	.24	.01
United Kingdom	-.31	.22	.38
United States	-.24	.01	.58

Standardised LISREL estimates. Based on OECD calibration sub-samples.

Table 4 shows the correlation between the latent variables as estimated by LISREL across country sub-samples, these estimates are reflecting the correlation between the constructs as measured without error. Whereas between DISCI and TSUP there is a moderate negative correlation in most countries, there is more variation in the correlation between ACHPR and the other two scales.

These findings do not necessarily imply lack of construct validity, as differences in relationship between these kinds of constructs may be due to differences in the instructional context. However, in view of the already apparent weaknesses of the Achievement Press construct, these inconsistent results may be interpreted as an additional indication of questionable construct validity.

For PISA 2000, these items were scaled using the IRT Partial Credit Model. Item fit was reviewed on the international level only. For future validation of indices it is planned to also test parameter invariance and review item fit after constraining the item parameters. This will give evidence to what extent the relative 'item difficulty'[4] is similar across countries. An additional review of step parameters might also be considered, but exisiting evidence shows that such a ,strong' validation typically leads to the rejection of international scaling models as it requires very similar response patterns across countries and increases the likelihood of finding discrepancies even further (Wilson, 1994).

4 That is, the extent to which an item is easier or harder to agree with respect to the latent trait.

EXAMPLE 2: FAMILY WEALTH

One important focus of the PISA 2000 study was the extensive coverage of student family background. In addition to family structure, parental education (ISCED classification), occupational status, language use at home, and immigrant background, students were asked about the existence of household possessions.

As Buchmann (2000) notes, collecting household possessions as indicators of family wealth has received much attention in recent international studies in the field of education. Household assets are believed to capture wealth better than income because they reflect a more stable source of wealth. TIMSS also used household possessions to measure home background, allowing countries to include a list of specific indicators together with an international core set of items.

Three different indices were derived from a set of indicators used in all participating PISA countries: Cultural Possessions, Home Educational Resources and Family Wealth (see Schulz, 2002). The dimensional structure was largely confirmed through exploratory factor analysis and a confirmatory factor analysis based on the OECD calibration sample. However, the dimensional structure varied considerably across countries and there was evidence that many of these indicators had somewhat different meanings in different cultural contexts.

The data used here were collected in the 32 participating countries in PISA 2000 and additional data from 10 non-OECD countries participating in the 'PISA plus' Study of 2001. As the country results for 'PISA plus' have not been released yet, their results will be included without any country identification.

Table 5: Items used to measure Family Wealth (PISA 2000 and PISAplus)

Q21		**In your home, do you have:**
	a)	a dishwasher
	b)	a room of your own
	c)	educational software
	d)	a link to the Internet
Q22		**How many of these do you have at your home? ('none', '1', '2', '3 or more')**
	a)	<Cellular> phone
	b)	Television
	d)	Computer
	f)	Motor car
	g)	Bathroom

Table 5 contains numbers and wording of the nine items used to measure Family Wealth. Four of these items (21a to d) were dichotomous (yes/no), the other five items categorical (Q22a, b, d, f and g). From the list of items it becomes apparent that IT-related items might as well be regarded as indicators of a family with interests in computers and factor analyses indeed showed evidence of an overlapping ‚IT factor'

explaining additional variance. However, all of these items are also indicators of family wealth because they reflect means of acquisition.

As to be expected, the percentages are very different across countries, probably depending on the socio-economic structure, development and wealth of a country. For example, whereas in some industrialised countries the percentage of students reporting to have a dishwasher at home is close to 90 percent, it is below 10 percent in some developing and Eastern European countries. In developing countries over 80 percent of students report not to have any computer at home, in some developed countries around 50 percent of students responded to have more than 2 computers in the home.

Tables 6 and 7 show the item-score correlations for the nine items used to measure family wealth across countries ordered by the overall scale reliabilities for each country: Generally, having a dishwasher at home or a room of his/her own are weaker indicators than others, however, in some countries the item-total reliabilities for the ‚dishwasher' item are very high. It can be observed that many items have lower item-score correlations in developed countries and that more developed countries tend to have lower scale reliabilities.

This is probably due to the fact that assets like computers, cars, mobile phones are still relatively expensive and often hardly affordable for many families in developing countries. In industrialised countries, however, these kinds of items are accessable for a vast majority of families and it is quite common to have them at home.

Table 6: Item-Score Correlations across Countries Ordered by Scale Reliability

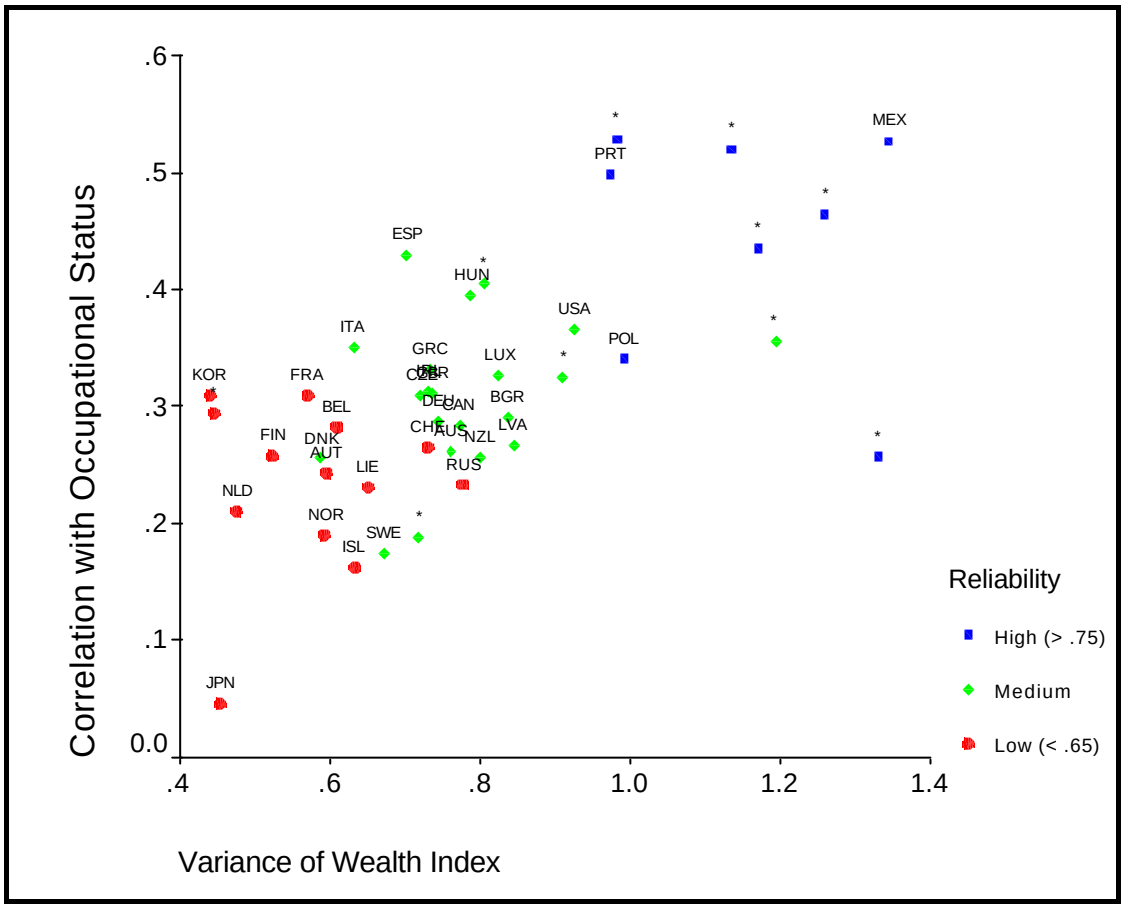
Country	Reliability	Dishwasher	Own room	Educational Software	Internet
Mexico	0.80	0.26	0.31	0.58	0.55
*	0.78	0.19	0.26	0.54	0.57
Brazil	0.78	0.36	0.15	0.61	0.60
*	0.77	0.22	0.62	0.59	0.13
*	0.77	0.05	0.20	0.47	0.57
Poland	0.76	0.41	0.26	0.52	0.47
*	0.75	0.44	0.34	0.24	0.53
Portugal	0.75	0.47	0.17	0.53	0.47
*	0.73	0.39	NA	0.18	0.46
*	0.72	NA	0.20	0.49	0.28
United States	0.72	0.45	0.25	0.46	0.52
*	0.70	0.23	0.17	0.54	0.50
Greece	0.70	0.37	0.16	0.40	0.47
Hungary	0.70	0.14	0.18	0.46	0.35
Canada	0.69	0.39	0.13	0.39	0.42
*	0.68	0.29	0.25	0.40	0.27
*	0.68	0.12	0.15	0.23	0.24
Latvia	0.68	0.25	0.17	0.47	0.38
New Zealand	0.68	0.40	0.15	0.39	0.44
Sweden	0.68	0.40	0.19	0.32	0.38
Czech Republic	0.67	0.35	0.19	0.42	0.39
Spain	0.67	0.40	0.11	0.37	0.41
Italy	0.67	0.37	0.18	0.36	0.43
Luxembourg	0.67	0.32	0.21	0.32	0.44
Australia	0.66	0.39	0.13	0.29	0.39
Denmark	0.66	0.37	0.17	0.27	0.38
Ireland	0.66	0.43	0.16	0.44	0.42
Germany	0.65	0.36	0.22	0.27	0.39
United Kingdom	0.65	0.41	0.13	0.40	0.42
Iceland	0.64	0.29	0.12	0.25	0.29
Switzerland	0.63	0.30	0.16	0.25	0.37
Russia	0.63	0.22	0.17	0.48	0.33
Austria	0.62	0.29	0.19	0.23	0.35
Finland	0.62	0.33	0.15	0.33	0.42
*	0.62	0.08	0.21	0.21	0.35
Belgium	0.61	0.31	0.09	0.28	0.38
France	0.61	0.33	0.18	0.37	0.36
Korea	0.61	0.20	0.22	0.34	0.38
Liechtenstein	0.60	0.25	0.12	0.28	0.38
Norway	0.60	0.25	0.16	0.24	0.30
Netherlands	0.53	0.22	0.13	0.24	0.33
Japan	0.50	0.15	0.16	0.21	0.30
Average	0.67	0.30	0.19	0.37	0.40

NA = indicator not included in questionnaire.

Table 7: Item-Score Correlations across Countries Ordered by Scale Reliability (continued)

Country	Reliability	Cellular phone	TV	Computer	Car	Bathroom
Mexico	0.80	0.60	0.56	0.65	0.63	0.53
*	0.78	0.51	0.48	0.66	0.53	0.52
Brazil	0.78	0.53	0.52	0.66	0.56	0.51
*	0.77	0.57	0.51	0.65	0.53	0.43
*	0.77	0.65	0.56	0.63	0.54	0.53
Poland	0.76	0.55	0.36	0.57	0.50	0.40
*	0.75	0.46	0.51	0.51	0.52	0.41
Portugal	0.75	0.49	0.38	0.61	0.50	0.42
*	0.73	0.56	0.40	0.55	0.50	0.41
*	0.72	0.58	0.46	0.53	0.43	0.49
United States	0.72	0.41	0.23	0.51	0.38	0.53
*	0.70	0.50	0.32	0.53	0.44	0.25
Greece	0.70	0.41	0.31	0.53	0.40	0.40
Hungary	0.70	0.50	0.34	0.52	0.53	0.37
Canada	0.69	0.43	0.27	0.45	0.40	0.49
*	0.68	0.46	0.43	0.42	0.46	0.33
*	0.68	0.56	0.47	0.49	0.46	0.51
Latvia	0.68	0.50	0.35	0.48	0.44	0.33
New Zealand	0.68	0.41	0.30	0.46	0.38	0.41
Sweden	0.68	0.44	0.37	0.43	0.39	0.36
Czech Republic	0.67	0.41	0.25	0.50	0.44	0.28
Spain	0.67	0.36	0.28	0.50	0.39	0.40
Italy	0.67	0.33	0.29	0.47	0.39	0.34
Luxembourg	0.67	0.41	0.33	0.49	0.41	0.34
Australia	0.66	0.42	0.29	0.43	0.35	0.42
Denmark	0.66	0.39	0.29	0.45	0.38	0.36
Ireland	0.66	0.28	0.24	0.46	0.38	0.40
Germany	0.65	0.32	0.32	0.46	0.43	0.35
United Kingdom	0.65	0.30	0.23	0.38	0.45	0.37
Iceland	0.64	0.42	0.41	0.37	0.41	0.32
Switzerland	0.63	0.37	0.31	0.40	0.38	0.33
Russia	0.63	0.37	0.34	0.48	0.37	0.25
Austria	0.62	0.26	0.35	0.41	0.38	0.30
Finland	0.62	0.36	0.25	0.42	0.31	0.24
*	0.62	0.34	0.38	0.44	0.38	0.40
Belgium	0.61	0.29	0.24	0.45	0.43	0.28
France	0.61	0.29	0.15	0.46	0.34	0.28
Korea	0.61	0.27	0.24	0.40	0.39	0.29
Liechtenstein	0.60	0.29	0.34	0.33	0.34	0.31
Norway	0.60	0.33	0.32	0.41	0.32	0.32
Netherlands	0.53	0.28	0.27	0.31	0.32	0.15
Japan	0.50	0.25	0.30	0.29	0.25	0.14
Average	**0.67**	**0.42**	**0.35**	**0.48**	**0.42**	**0.37**

Graph 2: Relationship between the variance of WEALTH and its correlation with Highest Occupational Status (HISEI)



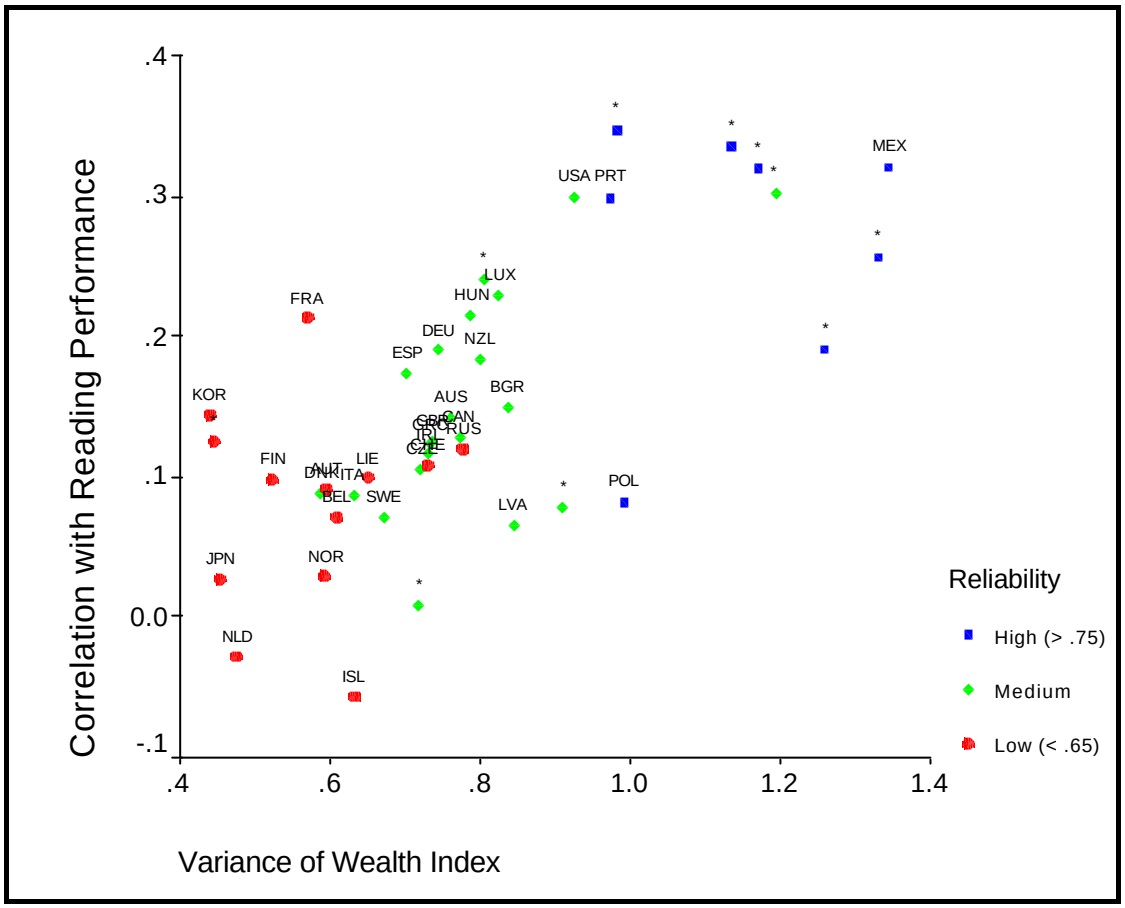
Dots represent locations of country values.

To assess the validity of the Wealth index it is also important to look at its relationship with other indices of socio-economic background. Graph 2 shows the relationship between correlations of the Wealth index with an index of occupational status of both parents (HISEI) [5] and the variance of this index. Additionally, countries are divided into three groups of those with high (Cronbach's Alpha > .75), medium and low (< .65) scale reliabilities and are marked accordingly.

The relationship between the variance in the Wealth index and its correlation with the occupational status of parents is fairly strong (r = .64). As can be observed there is also a relationship of the level of scale reliability with both variables: Countries with a low variance in the Wealth index tend to have a lower reliability and also show a weaker relationship with parental occupational status.

5 The mother's occupation and the father's occupation were observed through open-ended questions that were coded into the International Standard Classification of Occupation. These ISCO categories were then transformed into an International Socio-Economic Index according to the methodology developed by Ganzeboom, de Graaf and Treiman (1992). HISEI corresponds to the higher value of the mother's and the father's occupational status.

Graph 2: Relationship between the variance of WEALTH and its correlation with Reading Performance (1st PV)



Dots represent locations of country values.

A similar picture emerges when looking at the predictive power of Family Wealth for student performance in Reading (here, the first plausible value for the overall PISA Reading scale was used) and its relationship with the variance of this index. Graph 3 shows that the correlation of the Wealth index with student performance is also related to its within-country variance (r = .69). Again, countries with high reliability have higher levels of variance and stronger related to student performance.

Typically, the wealth index appears to be stronger in developing countries, in terms of scale reliability, its relationship with other background variables and its predictive power. This is particular the case in Latin American countries. However, it should be noted that in the United States the Wealth indicator is quite strongly related to other background measures. There is evidence that the indicators used in PISA 2000 were appropriate for societies - particularly from the Western Hemisphere - with larger gaps between low- and high-income families. Furthermore, not all of these indicators are equally adequate to measure family wealth in all participating countries.

Consequently, there is a need for a broader set of more country-specific indicators that provide more reliable and also more valid measures of family wealth. IRT methodology would enable PISA to scale a set of core items (with constrained item

parameters) together with country-specific items (with unconstrained item parameters), and still achieve measures that are comparable across countries.

DISCUSSION

In international studies there is a need for comparable measures about student background. Whereas in the past a vast amount of item analysis and an increasingly sophisticated scaling methodology has been spent on safeguarding the comparability and validity of cognitive tests, less attention has been paid to validity problems of indices derived from context questionnaire data. As the aim of international educational studies is not only to compare student performance across countries, but also to explain differences with student background measures, instructional context and school organisation, the credibility of international educational research rests on the comparability of these kinds of measures. Consequently, there is a need for investing more time and efforts in this field.

Structural Equation Modelling can be employed to validate questionnaire constructs internationally and provides a useful tool for reviewing item dimensionality and model fit across countries. However, using IRT for an analysis of country-specific item properties would be a more rigid test of parameter invariance across countries. But there is evidence that a strong construct validation requires a high amount of similarity in response patterns and that this approach might very often lead to the rejection of questionnaire constructs. So it needs to be discussed to what extent discrepancies in scaling models can be tolerated and what is deemed to be an adequate cross-country validation for this kind of measures.

Often, as in the case of the PISA index of Family Wealth, only a limited set of indicators is available for the measurement of important dimensions. Household possessions are an efficient way of collecting information about home background because they typically attract less missing responses than for example questions about parental education. The disadvantage of using household items as indicators of student background is that the meaning of these indicators can vary considerably across countries, depending on general income levels, income distribution and other country-specific factors.

The analysis of PISA data reveals that the appropriateness of the indicators depends largely on the national context. In some participating countries the index seems to work well, in others this measure might be improved through the inclusion of country-specific items that are more appropriate for the respective national context. Here, IRT methodology provides a tool to scale a set of international core items with constrained parameters together with additional national household items with unconstrained parameters.

REFERENCES

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.

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.

Buchmann, C. (2000). Measuring Family Background in International Studies of Educational Achievement: Conceptual Issues and Methodological Challenges. Paper presented at a symposium convened by the Board on International Comparative Studies in Education of the National Academy of Sciences/National Research Council on November 1, 2000, in Washington, D.C.

Ganzeboom, H.B.G., de Graaf, P.M., and Treiman, D.J. (1992). A standard international socio-economic index of occupational status. *Social Science Research*, *21*, 1-56.

Grisay, A. (2002). Translation and Cultural Appropriateness of the Test and Survey Material. In: Adams, R. and Wu, M. (Ed.). *Technical Report for the OECD Programme for International Student Assessment*, Paris: OECD Publications, pp. 57-70.

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

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.

OECD (2001). Knowledge and Skills for Life. First Results from PISA 2000. Paris: OECD.

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*, Paris: OECD Publications, pp. 217-252.

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. Norwood, NJ: Ablex, pp. 271-292.

APPENDIX

Table A: Item Reliabilities for Disciplinary Climate (DISCI)

	Disciplinary Climate					
	e	f	g	h	i	j
Australia	43	34	52	57	60	48
Austria	58	18	33	52	66	42
Belgium	53	21	48	41	56	43
Canada	49	32	48	49	63	45
Czech Republic	47	20	46	42	53	31
Denmark	39	28	43	49	58	29
Finland	60	18	49	61	67	51
France	51	25	31	33	68	29
Germany	49	19	38	49	58	41
Greece	45	27	24	16	61	13
Hungary	49	49	52	47	53	20
Iceland	49	27	30	51	63	44
Ireland	50	23	55	69	62	47
Italy	53	23	51	61	61	27
Japan	22	30	64	63	54	44
Korea	10	34	45	48	51	51
Mexico	26	18	26	43	48	36
Netherlands	52	33	51	59	63	40
New Zealand	36	28	32	55	74	42
Norway	43	41	49	59	56	38
Poland	45	22	37	54	61	45
Portugal	36	11	21	39	63	44
Spain	50	21	39	59	58	37
Sweden	37	34	38	52	57	24
Switzerland	50	7	27	44	59	51
United Kingdom	53	31	53	63	59	43
United States	34	42	61	62	59	35

Item reliabilities are derived from the squared standardised item loading multiplied with 100.

Table B: Item Reliabilities for Teacher Support (TSUP)

	Teacher Support					
	a	l	m	n	p	q
Australia	50	37	57	65	75	71
Austria	38	42	59	62	76	53
Belgium	40	42	55	52	66	43
Canada	50	44	60	63	70	69
Czech Republic	22	30	39	35	62	38
Denmark	46	39	38	58	63	62
Finland	40	33	53	56	78	73
France	40	46	61	60	70	56
Germany	39	33	49	60	76	52
Greece	47	48	39	62	69	67
Hungary	41	39	68	62	69	48
Iceland	32	44	59	63	70	59
Ireland	57	46	59	55	79	66
Italy	28	44	58	48	67	57
Japan	40	45	68	63	82	74
Korea	32	36	19	52	65	58
Mexico	46	36	24	47	55	66
Netherlands	30	37	58	53	69	46
New Zealand	35	36	46	59	77	61
Norway	48	41	53	56	77	68
Poland	40	29	58	62	72	62
Portugal	43	44	50	52	66	68
Spain	38	39	58	64	74	71
Sweden	54	47	50	55	81	64
Switzerland	36	34	63	61	79	49
United Kingdom	51	46	61	66	67	63
United States	55	43	55	67	73	76

Item reliabilities are derived from the squared standardised item loading multiplied with 100.

Table C: Item Reliabilities for Achievement Press (ACHPR)

	Achievement Press			
	c	d	e	o
Australia	20	18	32	16
Austria	21	40	23	6
Belgium	38	17	15	24
Canada	25	28	26	15
Czech Republic	52	39	20	28
Denmark	37	0	6	22
Finland	44	29	27	22
France	35	51	30	20
Germany	13	49	29	6
Greece	21	44	17	2
Hungary	74	6	7	24
Iceland	39	23	24	27
Ireland	18	25	27	11
Italy	23	41	21	6
Japan	54	10	0	11
Korea	7	81	1	0
Mexico	26	42	4	16
Netherlands	23	44	8	9
New Zealand	26	46	33	12
Norway	52	14	13	13
Poland	46	12	9	49
Portugal	21	26	13	22
Spain	38	42	18	18
Sweden	39	19	16	37
Switzerland	24	39	36	16
United Kingdom	16	23	7	12
United States	41	21	15	22

Item reliabilities are derived from the squared standardised item loading multiplied with 100.