

ESTIMATING PISA STUDENTS ON THE IALS PROSE LITERACY SCALE

Kentaro Yamamoto, Educational Testing Service (Summer, 2002)

Before PISA, the International Adult Literacy Survey (IALS) was conducted multiple times over several years in many countries to assess literacy skills of those in the population older than 16 years of age. Citing the results of the IALS data, many contend that a strong and increasing correlation exists between skills and opportunity for individuals. This growing recognition of the important role that literacy skills play in providing individuals with the ability to fully participate in increasingly complex societies led PISA to measure the knowledge and skills that 15-year-old students need beyond the school doors.

With this in mind, a subset of the literacy tasks from the IALS prose literacy scale was embedded into the PISA assessment. The intention was two-fold. First, these items were coded as part of the PISA Reading Literacy framework and were included, along with items specifically constructed for use in PISA, to estimate the Reading literacy proficiencies of 15-year-old students in participating countries. Second, the IALS items were included in the PISA assessment to enable an estimation of the IALS Prose literacy proficiency of 15-year-olds in participating countries.

The results from the IALS are reported on three scales—Prose, Document and Numeracy—which the National Adult Literacy Survey (NALS) of the United States were reported in 1993. Along with the measurement construct of three literacy scales, the data collection method and data analysis method were modeled after those used for NALS. With scaling methods, the performance of a sample of respondents can be summarized on a series of scales even when different respondents have been administered different assessment items. The two major goals of the PISA/IALS linking study were to report on common scales the results of many different countries with different languages, and to relate the scale proficiency distributions of PISA reading scale with the IALS literacy scales. This section describes the models and procedures used to evaluate the item parameters of the IALS results, to estimate respondents' Prose literacy proficiencies for each country, and to conduct statistical analyses.

OVERVIEW OF ANALYSIS

PISA gathered background as well as cognitive item information on 174,923 sampled students according to a stratified sampling method prescribed by the survey's directors. Each student received a booklet containing some subsets of reading, mathematics, and science PISA items as well as prose literacy tasks of IALS. About 30% of the total sample group received some of IALS Prose scale literacy items and the breakdown is available in the table below.

Sample sizes by Country

Country	Total sample size	Sample size receiving IALS Prose Items
Australia	5176	1658
Austria	4745	1557
Belgium	6670	2063
Brazil	4893	1351
Canada	29687	9680
Czech Republic	5365	1671
Denmark	4235	1319
Finland	4864	1604
France	4673	1489
Germany	5073	1597
Greece	4672	1420
Hungary	4887	1478
Iceland	3372	1069
Ireland	3854	1247
Italy	4984	1566
Japan	5256	1675
Korea	4982	1644
Liechtenstein	314	99
Luxembourg	3528	1033
Mexico	4600	1443
Netherlands	2503	819
New Zealand	3667	1199
Norway	4147	1318
Poland	3654	1140
Portugal	4585	1481
Russian Federation	6701	2022
Spain	6214	1941
Sweden	4416	1401
Switzerland	6100	1875
United Kingdom	9340	3031
United States	3846	1234

Latvia	3920	1188
Total	174923	55312

Issues involved in linking PISA and IALS include establishing comparability between countries and between the PISA and IALS Prose scale scores. Comparability is supported by equivalencies of measurement instruments, administration, scoring, and analysis. For PISA, 15 Prose items were selected from the IALS item pool. These items were selected from items previously piloted in 21 countries, and results were used to further refine translation, scoring, and printing. The selected items were then divided into two sets of items, one set containing eight items and one set containing seven. These sets of items were embedded separately into six PISA booklets in varying positions in order to minimize the potential position effects. PISA managers verified scoring and rescoring reliabilities, and ETS received the verified scores for further analyses.

In addition to scoring reliability, the treatment of missing responses may introduce errors into the scaling procedure due to misattribution of the causes of not responding to items. In this study, as in IALS, all responses appearing before the last legitimate responses were treated as omitted, while all missing responses occurring after the last response were treated as not reached. This results in omitted items being treated as incorrect in cases where a respondent had the chance to produce a response and chose not to, and not reached items being treated as if they were not administered. From a scaling perspective, not reached items are not used in estimating proficiency, while omitted items are included as part of the estimation. It should be noted that this treatment is identical to the methods used in PISA (while item parameters are estimated but different from when students' proficiencies are generated. In the PISA study, the interpretation of not reached responses are treated "unrelated to ability" during the item calibration, but when proficiencies are estimated then not-reached responses are treated as incorrect. In order to maintain the strongest possible linkage between the PISA/IALS and the IALS scale, the scoring procedures of IALS were kept intact.

Block Position Effect

A block position effect, which is the interaction between item position in a booklet and performance, was evaluated. This concern is quite common in the field of Psychometrics and not limited to this survey. The booklet design of PISA was made in order to counter balance such possibility of position effect on the difficulty of items. There were nine booklets each consisting of four 30 minute blocks. There was a clear time break between the first two blocks and the last

two blocks. In terms of block position effect, administration condition would resemble to repeating two block sets. Two groups of Literacy items were included in two PISA blocks and appeared exactly three times in different positions. Block A appeared in booklets 1, 2, and 6, and block B appeared in booklets 7, 8, and 9. The position of blocks in the booklets is presented in the table below. Booklets 1, 2, and 6 included first three blocks of reading and followed by Mathematics and/or Science blocks in the end. All four blocks of booklet 7 were reading blocks. Booklet 8 and 9 had two reading blocks in the second set only.

Position of Literacy Item Blocks in PISA Booklets

Booklet	Set1		Break	Set2	
	First	Second		First	Second
1		Block A			
2	Block A				
6				Block A	
7					Block B
8				Block B	
9					Block B

Listed in the appendix is the averaged proportion of not-reached responses among those who received IALS literacy items. It is clear that more non-responses are present in second position and also in the set 2. In order to minimize the impact of not reached responses in the cognitive data analysis all subsequent data analysis used only three response categories of incorrect, correct and omit responses

We applied the analysis of variance method to evaluate possible order effect. We found the F statistics were significant for both blocks, thus there is strong evidence for the block position effect. The following table shows the proportions correct at block level averaged across all countries.

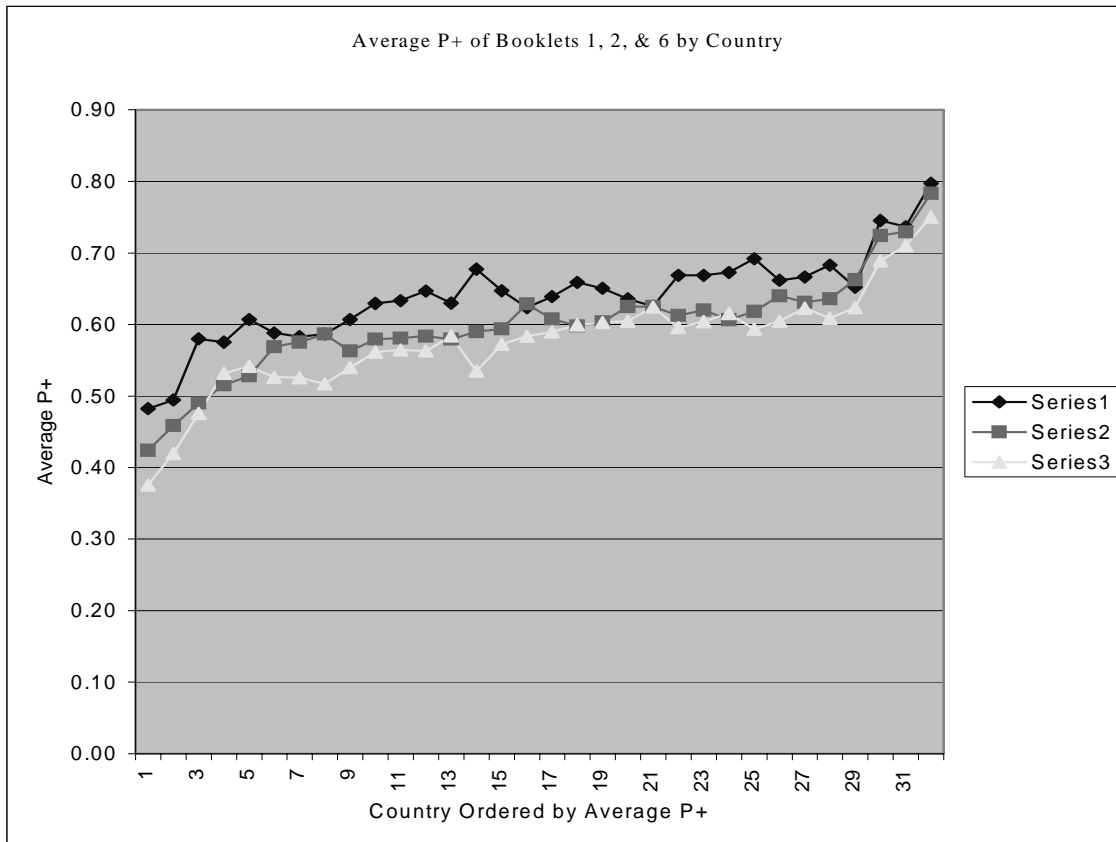
Weighted proportion correct of an item was calculated as follows,

$$P = \frac{\sum correct}{\sum correct + \sum wrong + \sum omit}$$

Note: The denominator represents the total attempts. Not-reached responses, consecutively missing responses at the end of a block, were not included in calculating proportions correct. Exclusion of not-reached items from proportions correct is consistent with scaling procedures applied to produce proficiency scores based on the IRT model for IALS as well as for PISA/IALS.

Average Percent Correct; Block by Block Order

Block Order	Block A	Block B
First	63.9	51.0
Second	57.7	49.7
Last	59.9	45.8
Position effect F.	64.5(0.0001)	108.7(0.0001)



Block position effects were evident for nearly every country for both blocks, while block position effects were never found in the IALS study (IALS technical report). It is not certain why a block position effect for PISA occurs even after not reached responses have been eliminated. It may be due to PISA being a longer test than the IALS and/or because PISA is classroom-administered under time-limited conditions, unlike the IALS assessment, which was administered at home without practical time constraints.

Scaling Methodology

This section reviews the scaling model employed in the analyses of the PISA/IALS data and explains the multiple imputation or “plausible values” methodology.

The scaling model

The scaling model used for the PISA/IALS is the two-parameter logistic (2PL) model from item response theory (Birnbaum, 1968; Lord, 1980). It is a mathematical model for the probability that a particular person will respond correctly to a particular item from a single domain of items. This probability is given as a function of a parameter characterizing the proficiency of that person, and two parameters characterizing the properties of that item. The following 2PL IRT model was employed in the IALS:

$$P(x_{ij} = 1 | \theta_j, a_i, b_i) = \frac{1}{1.0 + \exp(-Da_i(\theta_j - b_i))} \quad (1)$$

where

- x_{ij} is the response of person j to item i , 1 if correct and 0 if incorrect;
- θ_j is the proficiency of person j (note that a person with higher proficiency has a greater probability of responding correctly);
- a_i is the slope parameter of item i , characterizing its sensitivity to proficiency;
- b_i is its location parameter, characterizing its difficulty.

Note that this is a monotone increasing function with respect to θ ; that is, the conditional probability of a correct response increases as the value of θ increases. In addition, a linear indeterminacy exists with respect to the values of θ_j , a_i , and b_i for a scale defined under the two-parameter model. In other words, for an arbitrary linear transformation of θ say $\theta^* = M\theta + X$, the corresponding transformations $a_i^* = a_i/M$ and $b_i^* = Mb_i + X$ give:

$$P(x_{ij} = 1 | \theta_j^*, a_i^*, b_i^*) = P(x_{ij} = 1 | \theta_j, a_i, b_i)$$

Upon validation of the IALS scales with the PISA samples, the same linear transformation of the scales used for the IALS can be used.

Another main assumption of IRT is conditional independence. In other words, item response probabilities depend only on θ (a measure of proficiency) and the specified item parameters, and not on any demographic characteristics of student, or on any other items presented together in a test, or on the survey administration conditions. This enables us to formulate the following joint probability of a particular response pattern x across a set of n items.

$$P(\underline{x}|\theta, \underline{a}, \underline{b}) = \prod_{i=1}^n P_i(\theta)^{x_i} (1 - P_i(\theta))^{1-x_i}$$

Replacing the hypothetical response pattern with the real scored data, the above function can be viewed as a likelihood function that is to be maximized with a given set of item parameters. These item parameters were treated as known for the subsequent analyses.

Another assumption of the model is unidimensionality — that is, performance on a set of items is accounted for by a single unidimensional variable. Although this assumption may be too strong, the use of the model is motivated by the need to summarize overall performance parsimoniously within a single domain. Hence, item parameters were estimated for each scale separately.

Testing the assumptions of the IRT model, especially the assumption of conditional independence, is a critical part of the data analyses. The conditional independence means that respondents with identical abilities have a similar probability of producing a correct response on an item regardless of their group membership. In this study, country of residence was used as a grouping variable to test the conditional independence. This assumption applies to those subsamples within a country who received different sets of items. Serious violation of the conditional independence assumption would undermine the accuracy and integrity of the results. It is a common practice to expect a portion of items to be found not suitable for a particular subpopulation. Thus, while the item parameters were being estimated, empirical conditional percentages correct were monitored across the samples.

IALS Item parameter evaluation

Scale linking and Item parameter estimation

In IALS, the estimation of item-parameters for the 2PL model has been carried out using a modified version of Mislevy and Bock's (1982) BILOG program. BILOG procedures are based on an extension of the marginal-maximum-likelihood approach described by Bock and Aitkin (1981). The IALS version of BILOG maximizes the likelihood

$$L(\beta) = \prod_g \prod_{i,g} \int_{\theta} P(x_{i,g} | \theta, \beta) f_g(\theta) d(\theta)$$

$$\approx \prod_g \prod_{i,g} \sum_k P(x_{i,g} | \theta = X_k, \beta) A_g(X_k)$$

In the equation, $P(x_{j,g} | \theta, \beta)$ is the conditional probability of observing a response vector x_{jg} of person j from group g , given proficiency θ and vector of item parameters $\beta = (a_1, b_1, \dots, a_i, b_i)$ and $f_g(\theta)$ is a population density for θ in group g . Prior distributions on item parameters can be specified and used to obtain Bayes-modal estimates of these parameters (Mislevy, 1984). The proficiency densities can be assumed known and held fixed during item parameter estimation or estimated concurrently with item parameters.

The f_g in the above equation are approximated by multinomial distributions over a finite number of “quadrature” points, where X_k , for $k=1, \dots, q$, denotes the set of points and $A_g(X_k)$ are the multinomial probabilities at the corresponding points that approximate $f_g(\theta)$ at $\theta=X_k$. If the data are from a single population with an assumed normal distribution, Gauss-Hermite quadrature procedures provide an “optimal” set of points and weights to best approximate the integral for a broad class of smooth functions. For more general f or for data from multiple populations with known densities, other sets of points (e.g., equally spaced points) can be substituted and the values of $A_g(X_k)$ may be chosen to be the normalized density at point X_k (i.e., $A_g(X_k) = f_g(X_k) / \sum_k f_g(X_k)$).

Maximization of $L(\beta)$ is carried out by an application of an EM algorithm (Dempster, Laird, & Rubin, 1977). When population densities are assumed known and held constant during estimation, the algorithm proceeds as follows. In the E-step, provisional estimates of item parameters and the assumed multinomial probabilities are used to estimate “expected sample sizes,” at each quadrature point for each group, $\hat{N}_{g,k}$. These same provisional estimates are also used to estimate an “expected frequency” of correct responses at each quadrat point for each group, $\hat{r}_{g,k}$. In the M-step, improved estimates of the item parameters are obtained by treating the $\hat{N}_{g,k}$ and $\hat{r}_{g,k}$ as known and carrying out maximum-likelihood logistic regression analysis to estimate the item parameters β , subject to any constraints associated with prior distributions specified for β .

Using a current version of Yamamoto's (1989) HYBIL computer program, the two-parameter fit of logistic IRT model item parameters was evaluated using sample weights for each country separately. Original procedures used to estimate IALS item parameters were described in the technical report (1998).

The IALS cognitive items were estimated in 1994 using over 25,000 adults and further validated in 1996 and then again in 1998.

The IALS item parameters calibrated for the PISA study must fit well in order to justify the use of the item parameter estimates without modification. A graphical method as well as a χ^2 statistic was used to verify such fit. The statistic indicated a very good fit.

Standardized sample weights were used during item calibration. It is known that different subpopulation distributions occur within different assessment samples. Such variations may arise because of differences in the characteristics of the target populations, the sampling design, or the randomness of sampling. By applying post-stratified weights, vital characteristics of the sample can be closely matched to the characteristics of the population. During calibration, the fit of item parameters is maximized about the proficiency distribution of the calibration sample. When item parameters are being estimated, it is ideal to match the proficiency distribution of the calibration sample as closely as possible to that of the population. It is more critical when item calibration is done on the combined proficiency distribution of multiple assessment samples with great differences in proficiency distributions, such as the PISA. Standardization took place among countries as well. If we do not consider differences in national populations, using sampling weights representing population in each country would result reflecting primarily the size of population in each country. This would impact on large countries dominating the item parameter calibration, and may not be optimal for some smaller countries. In order to ensure the fit of item parameters to be comparable across countries, the total sum of standardized population weights in each country was set to a constant.

To obtain unbiased parameter estimates, proficiency distributions for the separate assessment samples were estimated during calibration. Each country received separate empirical prior normal proficiency distribution according to the response probability, and they were updated every iteration. It is known that the samples for each assessment came from somewhat different populations with different characteristics. The calibration procedure should take into account the possibility of systematic interaction of samples and items to estimate unbiased

estimates of sample distributions and item parameters. For that reason, a normal distribution with a unique mean and variance for the population of each country was estimated concurrently with item parameters. The reason for not using multinomial distribution for the prior distribution is its inherent instability of the shape of the distribution due to the over parameterization.

There are two options for accommodating the misfit of the IRT model while keeping the international scales intact. One approach is to drop the deviant items from the analysis. A drawback of this option is that it results in a smaller number of items, especially if items are dropped when the IRT functions differ in only one or two countries. We would use this approach if the IRT model did not fit at all, for example, if the response function was negative, or if all observed response functions were so far apart from each other that one set of item parameters would not describe responses from most of the countries. The approach used in this study was to psychometrically model large deviations by estimating best fitting item parameters. If there were some systematic deviations among multiple countries, they shared the same item parameters.

A novel procedure to constrain item parameters across countries was used. For a few items, misfits of IALS item parameters were very similar across countries in terms of direction of deviation and its magnitude. Thus, a pair of item parameters was estimated based on a set of countries. Then for a particular item, in addition to the original IALS item parameters, another set of item parameters is applicable. This reduced number of item parameters to be estimated compared to estimating unique item parameters for every country that can be 960 altogether.

The IALS item parameters must fit well in order to justify the use of the item parameter estimates without modification. A graphical method as well as χ^2 statistics and square root of weighted Mean Squared Deviation, and weighted Mean Deviation were used to verify such fit at an item level for every country separately. Deviations are based on the difference between model-based expected proportions correct and observed proportions correct at each equally spaced 31 ability scale values. Upon examination of the fit statistics and indices, a certain regularity across many countries and irregularity unique to a particular country was evident. On average 4.4 items per country had absolute weighted mean deviation of greater than 0.1 in terms of proportions correct. Altogether 19 sets of item parameters (36 altogether) were estimated in addition to the 30 item parameters originally estimated using IALS data. There were potentially 960 ($32 \times 15 \times 2$) item parameters to account for 32 countries, and 15 items of 2 parameters per item when each country get unique sets of item parameters. Sixty-four population parameters account for the

mean and standard deviations of 32 countries, and must be estimated. It is clear that by estimating just 36 more item parameters the model fit greatly improved as evidenced in substantial reduction of $-2*\text{Log-Likelihood}$ of 974,132 to 935,814.

Model Fit of Original IALS Item Parameters and Final Estimated Parameters

Model estimation	-2*Log Likelihood	Average RMSD	Average MD	Average # items $ MD >0.1$	No. of item parameters	No. of pop. parameters
Original IALS parameters	974132	0.091	-0.0098	4.4	30	64
Final Parameters	935814	0.052	-0.0014	0.8	66	64

For a majority of the items, 10 out of 15, one pair of item parameters was estimated in addition to the original IALS item parameters. Also it was most common to see that item parameters were estimated for only one country out of 32. Four items indicated more systematic deviations from IALS. However, the comparability among all 32 countries is relatively intact due to common item parameters being estimated for these four items for the majority of countries.

Summary of Number of Item Parameters Estimated

	Original parameters	Original parameters +1 pair	Original parameters +2 pairs	Total No. of items
No. of items	1	10	4	15
No of parameters	2	40	24	66

A more detailed design of IRT parameter estimation representing how item parameters are constrained to support the common IALS scale is presented below through the levels of item parameter estimation. There were 15 items altogether. Level “0” represents the item parameters being identical to the IALS for the item by country pair. Level “1” represents the all 1s in the column were estimated together, so they are the same among themselves and they are different from the IALS parameters. Level “2” represents the all 2s in the column were estimated together and different from level 1 and IALS parameters. For three countries, item parameters were not estimated for one or two items due to these items being dropped by the PISA management before the IALS Linking analysis took place. An “X” indicates these country item pairs. Estimated item parameters for each country are listed in the Appendices.

Level of Non-IALS Parameters on Item by Country

Code#	Country	Items															Non IALS parameters
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
36	Greece							1	1	1	2			1			5
40	United Kingdom								1	1	1			1			4
56	France					1				1	1			1			4
76	United States						1			1	1						3
100	Latvia		1								1		2	1		1	5
124	Czech Republic						1		1		1						3
156	Luxembourg								1	1			1	1			4
203	Korea		1				X		1	1				1		X	4
208	Finland								1		1			1			3
246	Poland								1	1	1			1			4
250	Austria								1				1	1	1		4
276	Ireland								1	1	1			1			4
300	Belgium						1		1	1	1						4
348	Portugal								1	1	1			1			4
352	Australia								1	1	1						3
372	Liechtenstein				1				1		X						2
380	Italy			1					1		1	1		1			5
392	Canada								1	1	1						3
410	Netherlands						1	1	2	1	1						5
438	Norway								1	2	1		1	1			5
442	Spain								1	2	2			1			4
484	Iceland			X					1	1				1			3
528	Brazil							1	1	1	2						4
554	Mexico		1						1		1		2				4
578	Hungary									2	1		2	1			4
616	Switzerland						1		1	1	1						4
643	Sweden								1	2	1		1	1			5
724	Denmark								1	1	1			1			4
752	Japan		1					1	1	1				1			5
756	Russian Federation			1					1				2	1			4
826	Germany								1	1			1	1			4
840	New Zealand						1			1	1						3
Non-IALS parameters		0	4	2	1	1	6	4	27	24	24	1	9	21	1	1	129

“X” indicates the item parameters were not estimated due being dropped by the PISA management before the IALS Linking analysis took place.

PROFICIENCY ESTIMATION USING PLAUSIBLE VALUES

Generating Proficiency Scores

Most cognitive skills tests are concerned with accurately assessing the performance of individual respondents for the purposes of diagnosis, selection, or placement. Regardless of which measurement model is being used, classical test theory or item response theory, the accuracy of these measurements can be improved—that is, the amount of measurement error can be reduced—by increasing the number of items given to the individual. Thus, achievement tests containing more than 70 items are common. Since the uncertainty associated with each θ is negligible, the distribution of θ or the joint distribution of θ with other variables can be approximated using individual θ s.

When analyzing the distribution of proficiencies in a group of persons, however, more efficient estimates can be obtained from a sampling design similar to that was used in the IALS. The survey solicits relatively few responses from each sampled respondent while maintaining a wide range of content representation when responses are summed for all respondents. The advantage of estimating population characteristics more efficiently is offset by the inability to make precise statements about individuals. Uncertainty associated with individual θ estimates is too large to be ignored. Point estimates of proficiency that are, in some sense, optimal for each sampled respondent could lead to seriously biased estimates of population characteristics (Wingersky, Kaplan, & Beaton, 1987).

Plausible value methodology was developed as a way to estimate key population features consistently and to approximate others no worse than standard IRT procedures would. A detailed review of plausible value methodology is given in Mislevy (1991). Along with theoretical justifications, Mislevy presents comparisons with standard procedures, discusses biases that arise in some secondary analyses, and offers numerical examples.

The following is a brief survey of the plausible values approach, focusing on its implementation in the PISA analyses.

Let \underline{y} represent the responses of all samples respondents to background questions and questions on engagement to literacy activities, and let θ represent the scale proficiency values. If θ were known for all sampled examinees, it would be possible to compute a statistic $t(\theta, \underline{y})$ —such as a scale or composite subpopulation sample mean, a sample percentile point, or a sample regression coefficient—to estimate a corresponding population quantity T .

Because the scaling models are latent variable models, however, θ values are not observed even for sampled respondents. To overcome this problem, we follow Rubin (1987) by

considering θ as “missing data” and approximate $t(\theta, \underline{y})$ by its expectation given $(\underline{x}, \underline{y})$, the data that actually were observed, as follows:

$$\begin{aligned} t^*(\underline{x}, \underline{y}) &= E[t(\theta, \underline{y}) | \underline{x}, \underline{y}] \\ &= \int t(\theta, \underline{y}) p(\theta | \underline{x}, \underline{y}) d\theta \end{aligned}$$

It is possible to approximate t^* using random draws from the conditional distribution of the scale proficiencies given the item responses x_j , background variables y_j , and model parameters for sampled respondent j . These values are referred to as imputations in the sampling literature, and as plausible values in IALS. The value of θ for any respondent that would enter into the computation of t is thus replaced by a randomly selected value from his or her conditional distribution. Rubin (1987) proposed to repeat this process several times so that the uncertainty associated with imputation can be quantified by “multiple imputation.” For example, the average of multiple estimates of t , each computed from a different set of plausible values, is a numerical approximation of t^* of the above equation; the variance among them reflects uncertainty due to not observing θ . -It should be noted that this variance does not include the variability of sampling from the population.

It cannot be emphasized too strongly that plausible values are not test scores for individuals in the usual sense. Plausible values are only intermediary computations for calculating integrals of the form of the above equation in order to estimate population characteristics. When the underlying model is correctly specified, plausible values will provide consistent estimates of population characteristics, even though they are not generally unbiased estimates of the proficiencies of the individuals with whom they are associated. The key idea lies in a contrast between plausible values and the more familiar ability estimates of educational measurement that are in some sense optimal for each respondent (e.g., maximum likelihood estimates, which are consistent estimates of a respondent’s θ , and Bayes estimates, which provide minimum mean-squared errors with respect to a reference population). Point estimates that are optimal for individual respondents have distributions that can produce decidedly nonoptimal (inconsistent) estimates of population characteristics (Little & Rubin, 1983). Plausible values, on the other hand, are constructed explicitly to provide consistent estimates of population effects. For further discussion, see Mislevy, Beaton, Kaplan, and Sheehan (1992).

Plausible values for each respondent j are drawn from the conditional distribution $P(\theta_j | \underline{x}_j, \underline{y}_j, \Gamma, \sigma^2)$, where Γ is a matrix of regression coefficients and σ^2 is a common variance for

residuals. Using standard rules of probability, the conditional probability of proficiency can be represented as follows:

$$\begin{aligned}
 P(\underline{\theta}_j | \underline{x}_j, \underline{y}_j, \Gamma, \sigma^2) &\propto P(\underline{x}_j | \underline{\theta}_j, \underline{y}_j, \Gamma, \sigma^2) P(\underline{\theta}_j | \underline{y}_j, \Gamma, \sigma^2) \\
 &= P(\underline{x}_j | \underline{\theta}_j) P(\underline{\theta}_j | \underline{y}_j, \Gamma, \sigma^2)
 \end{aligned}
 \tag{2}$$

where θ_j is a vector of three scale values, $P(x_j|\theta_j)$ is the product over the scales of the independent likelihoods induced by responses to items, and $P(\theta_j|y_j, \Gamma, \sigma^2)$ is the density of proficiencies conditional on the observed value y_j of background responses and parameters Γ and σ^2 . Item parameters estimates are fixed and regarded as population values in the computation described in this section.

In the analyses of the PISA/IALS, a normal distribution was assumed for $P(\theta_j|y_j, \Gamma, \sigma^2)$, with a common variance, σ^2 , and with a mean given by a linear model with slope parameters, Γ , based on the first principal components of several hundred selected main effects of background variables. Two variables: age, and gender were conditioned directly. Based on the principal component method, components representing 80 percent of the variance present in remainder of the data were selected. The included principal components will be referred to as the conditioning variables, and denoted as y^c . (The complete set of original background variables used in the analyses are listed in the appendices.) The following model was fit to the data.

$$\theta = \Gamma' y^c + \varepsilon$$

where ε is normally distributed with mean zero and variance σ^2 . As in a regression analysis, Γ is a vector that is the effects and σ^2 is the variance of residuals.

For all functions Γ of θ to be accurate, it is necessary that $p(\theta|y)$ be correctly specified for all background variables in the survey. In the PISA, however, principal component scores based on the frequently contrast coded nearly all background variables were used. The computation of marginal means and percentile points of θ for these variables is nearly optimal. Estimates of functions T involving background variables not conditioned on in this manner are subject to estimation error due to misspecification. The nature of these errors was discussed in detail in Mislevy (1991), Thomas (2000), von Davier (2000). Their magnitudes diminish as each respondent provides more cognitive data—that is, responds to a greater number of items. Indications are that the magnitude of these errors is negligible in the IALS (e.g., biases in regression coefficients below 5 percent) due to the larger numbers of cognitive items presented to

each respondent in the survey (on average, 16 items per respondent per scale). However, PISA/IALS had on average 7.5 items per respondent.

The basic method for estimating Γ and σ^2 with the EM procedure was described in Mislevy (1985) for a single scale case. The EM algorithm requires the computation of the mean, θ and variance, σ_p^2 , of the posterior distribution in (2). The updated version of the computer program CGROUP (Thomas, 1993) was used. Similar to the item parameter estimation weights were employed in this step. Since each country was analyzed separately due to covariances of Γ and θ being unique to every country, the weights standardized within a country to sum up to the sample size were used.

After completing the EM algorithm, the plausible values are drawn in a three-step process from the joint distribution of the values of Γ for all sampled respondents with at least three cognitive items attempted. First, a value of Γ is drawn from a normal approximation to $P(\Gamma, \sigma^2 | x_j, y_j)$ that fixes σ^2 at the estimates (Thomas, 1993). Second, conditional on the generated value of Γ (and the fixed value of σ^2), the mean θ_{pj} , and variance σ_{pj}^2 of the posterior distribution in the equation (2) are computed using the same methods applied in the EM algorithm. In the third step, the θ_j are drawn from a multivariate normal distribution with mean θ_p and variance σ_{pj}^2 . These three steps are repeated five times, producing five imputations of θ_j for each sampled respondent.

For those with an insufficient number of responses, the Γ and σ^2 s described in the previous paragraph were fixed, then plausible values were drawn from the posterior distributions. Hence, all respondents—regardless of the number of items attempted—were assigned a set of plausible values. Note that the posterior variance of subjects with no or only a few cognitive responses is noticeably larger than those with complete cognitive item responses, so the uncertainty about the ability of these respondents will be reflected in a variability among the plausible values.

Evaluation of relationships of proficiency scores of PISA reading and IALS prose

Estimated PISA/IALS items are linked at the item level through calibration. However, after this linking, the IRT parameters are still on the provisional scale, i.e. not ready to be reported on the IALS Prose scales. The transformation constants must be applied to put the new provisional scale onto the IALS scales, and these constants are identical to those used for the IALS and are used as $PV = A * \theta + B$.

Transformation Constants Applied to Produce Reported Scale

Literacy scale	A	B
Prose	51.67	269.16

PISA reading items are not created following the same framework as IALS literacy items. This difference in the construct is expected to be seen in the relationship between the plausible values of the two assessments. Correlation between the two scale scores for entire population after correction for attenuation was 0.85, without correction the correlation was 0.73, based on those students who received both PISA reading items as well as IALS Prose items. This is very similar to the correlations between the IALS Prose, and Document literacy subscales was 0.89 after being corrected for attenuation. However, the correlation between the PISA and IALS scale scores is slightly overestimated due to the PISA reading scale scores including responses on the IALS Prose items. It was not possible to reanalyze for this purpose alone to evaluate more accurate relationships between the two scales.

References

- Beaton, A. E., & Johnson, E. G. (1990). The average response method of scaling. *Journal of Educational Statistics, 15*, 9-38.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley Publishing.
- Cohen (1988). *Statistical power analysis for the behavioral sciences*. (2nd ed.) Hillsdale, NJ: Lawrence Erlbaum Associates.
- Johnson, E. G., & Rust, K. F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics*.
- Little, R.J.A. & Rubin, D. B. (1983). On jointly estimating parameters and missing data. *American Statistician, 37*, 218-220.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum Associates.
- Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association, 80*, 993-97.
- Mislevy, R.J. (1990). Scaling procedures. In E.G. Johnson and R. Zwick, *Focusing the new design: the NAEP 1988 technical report* (No. 19-TR-20). Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service.
- Mislevy, R.J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177-196.
- Mislevy, R.J., Beaton, A., Kaplan, B.A., and Sheehan, K. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*(2), 133-161.
- Mislevy, R. J. & Bock, R. D. (1982). *BILOG: Item analysis and test scoring with binary logistic models* [Computer program]. Morresville, IN: Scientific Software.
- Mislevy, R. J. & Sheehan, K. (1987) Marginal estimation procedures. In A. E. Beaton, *Implementing the new design: The NAEP 1983-84 technical report* (pp. 293-360). (no. 15-TR-20) Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

Thomas, N. (1993). Asymptotic corrections for multivariate posterior moments with factored likelihood functions. *Journal of Computational and Graphical Statistics*, 2, 309-22.

Thomas, N. (2000). Assessing Model sensitivity of the imputation methods used in the National Assessment of Educational Progress. *Journal of Educational and Behavioral Statistics*, 25, 351-371.

Wingersky, M, Kaplan, B. A., & Beaton, A. E. (1987). Joint estimation procedures. In A. E. Beaton, *Implementing the new design: The NAEP 1983-84 technical report* (pp.285-92) (No. 15-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Yamamoto, K. & Muraki, E. (1991). *Non-linear transformation of IRT scale to account for the effect of non-normal ability distribution on item parameter estimation*. A paper presented at the annual 1991 American Educational Research Association meeting, Chicago, IL. 1991.

von Davier, M. (2002) A comparison of conditional and marginal direct estimation of group statistics. *Technical report* (No. RR-02-xx). Princeton, NJ: Educational Testing Service.

Average Percentage Not Reached at Block Position by Country

		Block A			Block B		
		Book2	Book6	Book1	Book8	Book7	Book9
1	Greece	0.1	0.5	5.8	1.3	14.2	23.6
2	United Kingdom	0.0	0.5	2.0	0.8	4.3	8.6
3	France	0.0	0.3	6.1	0.4	4.5	13.0
4	United States	0.3	1.9	5.4	1.1	1.8	9.5
5	Latvia	0.1	1.4	13.0	3.3	15.2	23.6
6	Czech Republic	0.0	0.0	5.6	0.1	2.8	11.0
7	Luxembourg	0.6	0.1	13.6	1.0	34.6	23.5
8	Korea	0.0	0.0	0.9	0.0	1.6	3.4
9	Finland	0.3	0.0	2.0	0.3	1.9	4.5
10	Poland	0.0	0.4	5.2	0.4	7.4	17.6
11	Austria	0.0	0.0	3.4	0.0	2.9	5.0
12	Ireland	1.4	0.2	3.0	0.2	1.9	8.4
13	Belgium	0.1	1.6	5.0	0.5	8.5	12.5
14	Portugal	0.2	0.4	7.2	0.7	4.6	11.6
15	Australia	0.0	0.1	2.6	0.7	3.0	6.2
16	Liechtenstein	0.0	0.0	7.1	0.0	6.0	25.3
17	Italy	0.0	0.6	9.6	1.1	7.0	17.1
18	Canada	0.6	0.4	2.8	0.8	2.6	7.0
19	Netherlands	0.0	0.0	0.6	0.0	1.4	2.3
20	Norway	0.2	1.8	2.6	2.1	5.0	13.8
21	Spain	1.3	0.7	9.4	0.9	5.1	17.8
22	Iceland	0.6	0.5	4.4	13.1	5.6	14.1
23	Brazil	0.3	2.0	29.0	3.5	21.1	37.7
24	Mexico	1.3	1.3	9.9	0.4	5.1	17.1
25	Hungary	0.1	0.2	12.0	1.9	4.6	16.0
26	Switzerland	0.0	0.0	6.1	9.1	11.7	19.0
27	Sweden	0.0	0.8	4.3	0.3	4.6	12.8
28	Denmark	0.0	0.6	9.1	0.2	7.7	19.2
29	Japan	0.1	0.4	6.0	1.2	6.7	11.5
30	Russian Federation	0.9	0.3	19.0	2.0	8.0	25.5
31	Germany	0.0	0.1	6.6	0.2	6.1	13.1
32	New Zealand	0.0	0.0	3.3	0.4	2.1	7.0
	Average	0.3	0.5	7.0	1.5	6.9	14.3

Average Percentage Correct, Incorrect and Omit Responses
at Block Level by Country

		Block A			Block B		
		correct	incorrect	omit	correct	incorrect	omit
1	Greece	54	37	9	50	27	23
2	United Kingdom	62	35	3	54	36	10
3	France	61	32	6	56	32	12
4	United States	59	38	3	51	42	7
5	Latvia	60	32	8	51	32	17
6	Czech Republic	56	38	6	52	35	13
7	Luxembourg	52	37	11	38	42	20
8	Korea	73	25	2	64	30	6
9	Finland	78	19	3	65	27	8
10	Poland	57	36	7	52	30	18
11	Austria	61	34	5	51	40	11
12	Ireland	63	34	3	59	34	7
13	Belgium	64	32	4	54	32	14
14	Portugal	56	37	7	41	44	15
15	Australia	63	34	3	56	35	9
16	Liechtenstein	63	32	6	51	27	22
17	Italy	63	29	7	52	33	15
18	Canada	64	33	3	60	33	7
19	Netherlands	65	33	2	54	42	4
20	Norway	60	35	6	46	38	16
21	Spain	61	34	5	46	40	14
22	Iceland	64	31	5	55	32	13
23	Brazil	43	41	16	38	39	23
24	Mexico	46	47	7	42	43	15
25	Hungary	59	35	6	49	36	15
26	Switzerland	62	31	7	48	37	15
27	Sweden	64	30	6	47	39	14
28	Denmark	56	35	9	52	31	17
29	Japan	72	23	5	62	23	15
30	Russian Federation	57	35	9	49	36	15
31	Germany	60	32	8	51	33	16
32	New Zealand	62	35	3	56	35	9
	Average	60.6	33.5	5.9	51.6	34.8	13.6

PISA/IALS Item Parameters 1

		B1Q10S1		B1Q11S1		B2Q6S1		B2Q7S1		B3Q7S1	
		A	B	A	B	A	B	A	B	A	B
1	Greece	1.167	-1.100	0.702	0.922	0.767	0.096	0.857	1.071	1.035	-0.073
2	United Kingdom	1.167	-1.100	0.702	0.922	0.767	0.096	0.857	1.071	1.035	-0.073
3	France	1.167	-1.100	0.702	0.922	0.767	0.096	0.857	1.071	1.072	-0.743
4	United States	1.167	-1.100	0.702	0.922	0.767	0.096	0.857	1.071	1.035	-0.073
5	Latvia	1.167	-1.100	0.774	-0.018	0.767	0.096	0.857	1.071	1.035	-0.073
6	Czech Republic	1.167	-1.100	0.702	0.922	0.767	0.096	0.857	1.071	1.035	-0.073
7	Luxembourg	1.167	-1.100	0.702	0.922	0.767	0.096	0.857	1.071	1.035	-0.073
8	Korea	1.167	-1.100	0.774	-0.018	0.767	0.096	0.857	1.071	1.035	-0.073
9	Finland	1.167	-1.100	0.702	0.922	0.767	0.096	0.857	1.071	1.035	-0.073
10	Poland	1.167	-1.100	0.702	0.922	0.767	0.096	0.857	1.071	1.035	-0.073
11	Austria	1.167	-1.100	0.702	0.922	0.767	0.096	0.857	1.071	1.035	-0.073
12	Ireland	1.167	-1.100	0.702	0.922	0.767	0.096	0.857	1.071	1.035	-0.073
13	Belgium	1.167	-1.100	0.702	0.922	0.767	0.096	0.857	1.071	1.035	-0.073
14	Portugal	1.167	-1.100	0.702	0.922	0.767	0.096	0.857	1.071	1.035	-0.073
15	Australia	1.167	-1.100	0.702	0.922	0.767	0.096	0.857	1.071	1.035	-0.073
16	Liechtenstein	1.167	-1.100	0.702	0.922	0.767	0.096	1.953	1.446	1.035	-0.073
17	Italy	1.167	-1.100	0.702	0.922	0.699	0.996	0.857	1.071	1.035	-0.073
18	Canada	1.167	-1.100	0.702	0.922	0.767	0.096	0.857	1.071	1.035	-0.073
19	Netherlands	1.167	-1.100	0.702	0.922	0.767	0.096	0.857	1.071	1.035	-0.073
20	Norway	1.167	-1.100	0.702	0.922	0.767	0.096	0.857	1.071	1.035	-0.073
21	Spain	1.167	-1.100	0.702	0.922	0.767	0.096	0.857	1.071	1.035	-0.073
22	Iceland	1.167	-1.100	0.702	0.922			0.857	1.071	1.035	-0.073
23	Brazil	1.167	-1.100	0.702	0.922	0.767	0.096	0.857	1.071	1.035	-0.073
24	Mexico	1.167	-1.100	0.774	-0.018	0.767	0.096	0.857	1.071	1.035	-0.073
25	Hungary	1.167	-1.100	0.702	0.922	0.767	0.096	0.857	1.071	1.035	-0.073
26	Switzerland	1.167	-1.100	0.702	0.922	0.767	0.096	0.857	1.071	1.035	-0.073
27	Sweden	1.167	-1.100	0.702	0.922	0.767	0.096	0.857	1.071	1.035	-0.073
28	Denmark	1.167	-1.100	0.702	0.922	0.767	0.096	0.857	1.071	1.035	-0.073
29	Japan	1.167	-1.100	0.774	-0.018	0.767	0.096	0.857	1.071	1.035	-0.073
30	Russian Federation	1.167	-1.100	0.702	0.922	0.699	0.996	0.857	1.071	1.035	-0.073
31	Germany	1.167	-1.100	0.702	0.922	0.767	0.096	0.857	1.071	1.035	-0.073
32	New Zealand	1.167	-1.100	0.702	0.922	0.767	0.096	0.857	1.071	1.035	-0.073

PISA/IALS Item Parameters 2

		B3Q9S1		B3Q11S1		B3Q12S1		B3Q13S1		B3Q15S1	
		A	B	A	B	A	B	A	B	A	B
1	Greece	1.248	-0.317	0.821	0.088	0.771	0.173	0.962	0.068	0.502	0.433
2	United Kingdom	1.248	-0.317	1.004	-0.576	0.771	0.173	0.962	0.068	0.628	-0.504
3	France	1.248	-0.317	1.004	-0.576	0.624	-0.354	0.962	0.068	0.628	-0.504
4	United States	1.157	0.121	1.004	-0.576	0.624	-0.354	0.962	0.068	0.628	-0.504
5	Latvia	1.248	-0.317	1.004	-0.576	0.624	-0.354	0.733	-0.584	0.628	-0.504
6	Czech Republic	1.157	0.121	1.004	-0.576	0.771	0.173	0.733	-0.584	0.628	-0.504
7	Luxembourg	1.248	-0.317	1.004	-0.576	0.771	0.173	0.962	0.068	0.719	-1.591
8	Korea			1.004	-0.576	0.771	0.173	0.962	0.068	0.719	-1.591
9	Finland	1.248	-0.317	1.004	-0.576	0.771	0.173	0.733	-0.584	0.628	-0.504
10	Poland	1.248	-0.317	1.004	-0.576	0.771	0.173	0.962	0.068	0.628	-0.504
11	Austria	1.248	-0.317	1.004	-0.576	0.771	0.173	0.733	-0.584	0.719	-1.591
12	Ireland	1.248	-0.317	1.004	-0.576	0.771	0.173	0.962	0.068	0.628	-0.504
13	Belgium	1.157	0.121	1.004	-0.576	0.771	0.173	0.962	0.068	0.628	-0.504
14	Portugal	1.248	-0.317	1.004	-0.576	0.771	0.173	0.962	0.068	0.628	-0.504
15	Australia	1.248	-0.317	1.004	-0.576	0.771	0.173	0.962	0.068	0.628	-0.504
16	Liechtenstein	1.248	-0.317	1.004	-0.576	0.771	0.173	0.733	-0.584		
17	Italy	1.248	-0.317	1.004	-0.576	0.771	0.173	0.733	-0.584	0.628	-0.504
18	Canada	1.248	-0.317	1.004	-0.576	0.771	0.173	0.962	0.068	0.628	-0.504
19	Netherlands	1.157	0.121	0.821	0.088	0.725	0.761	0.962	0.068	0.628	-0.504
20	Norway	1.248	-0.317	1.004	-0.576	0.771	0.173	0.656	0.772	0.628	-0.504
21	Spain	1.248	-0.317	1.004	-0.576	0.771	0.173	0.656	0.772	0.502	0.433
22	Iceland	1.248	-0.317	1.004	-0.576	0.771	0.173	0.962	0.068	0.719	-1.591
23	Brazil	1.248	-0.317	0.821	0.088	0.771	0.173	0.962	0.068	0.502	0.433
24	Mexico	1.248	-0.317	1.004	-0.576	0.771	0.173	0.733	-0.584	0.628	-0.504
25	Hungary	1.248	-0.317	1.004	-0.576	0.624	-0.354	0.656	0.772	0.628	-0.504
26	Switzerland	1.157	0.121	1.004	-0.576	0.771	0.173	0.962	0.068	0.628	-0.504
27	Sweden	1.248	-0.317	1.004	-0.576	0.771	0.173	0.656	0.772	0.628	-0.504
28	Denmark	1.248	-0.317	1.004	-0.576	0.771	0.173	0.962	0.068	0.628	-0.504
29	Japan	1.248	-0.317	0.821	0.088	0.771	0.173	0.962	0.068	0.719	-1.591
30	Russian Federation	1.248	-0.317	1.004	-0.576	0.771	0.173	0.733	-0.584	0.719	-1.591
31	Germany	1.248	-0.317	1.004	-0.576	0.771	0.173	0.962	0.068	0.719	-1.591
32	New Zealand	1.157	0.121	1.004	-0.576	0.624	-0.354	0.962	0.068	0.628	-0.504

PISA/IALS Item Parameters 3

		B4Q7S1		B6Q7S1		B6Q8S1		B7Q10S1		B7Q11S1	
		A	B	A	B	A	B	A	B	A	B
1	Greece	0.942	-0.157	1.132	-0.618	0.891	-0.679	1.417	-0.533	0.956	0.694
2	United Kingdom	0.942	-0.157	1.132	-0.618	0.891	-0.679	1.417	-0.533	0.956	0.694
3	France	0.942	-0.157	1.132	-0.618	0.891	-0.679	1.417	-0.533	0.956	0.694
4	United States	0.942	-0.157	1.132	-0.618	0.915	-0.106	1.417	-0.533	0.956	0.694
5	Latvia	0.942	-0.157	0.629	0.004	0.891	-0.679	1.417	-0.533	1.134	-0.128
6	Czech Republic	0.942	-0.157	1.132	-0.618	0.915	-0.106	1.417	-0.533	0.956	0.694
7	Luxembourg	0.942	-0.157	0.905	-1.359	0.891	-0.679	1.417	-0.533	0.956	0.694
8	Korea	0.942	-0.157	1.132	-0.618	0.891	-0.679	1.417	-0.533		
9	Finland	0.942	-0.157	1.132	-0.618	0.891	-0.679	1.417	-0.533	0.956	0.694
10	Poland	0.942	-0.157	1.132	-0.618	0.891	-0.679	1.417	-0.533	0.956	0.694
11	Austria	0.942	-0.157	0.905	-1.359	0.891	-0.679	1.162	-1.207	0.956	0.694
12	Ireland	0.942	-0.157	1.132	-0.618	0.891	-0.679	1.417	-0.533	0.956	0.694
13	Belgium	0.942	-0.157	1.132	-0.618	0.915	-0.106	1.417	-0.533	0.956	0.694
14	Portugal	0.942	-0.157	1.132	-0.618	0.891	-0.679	1.417	-0.533	0.956	0.694
15	Australia	0.942	-0.157	1.132	-0.618	0.915	-0.106	1.417	-0.533	0.956	0.694
16	Liechtenstein	0.942	-0.157	1.132	-0.618	0.915	-0.106	1.417	-0.533	0.956	0.694
17	Italy	0.748	-0.961	1.132	-0.618	0.891	-0.679	1.417	-0.533	0.956	0.694
18	Canada	0.942	-0.157	1.132	-0.618	0.915	-0.106	1.417	-0.533	0.956	0.694
19	Netherlands	0.942	-0.157	1.132	-0.618	0.915	-0.106	1.417	-0.533	0.956	0.694
20	Norway	0.942	-0.157	0.905	-1.359	0.891	-0.679	1.417	-0.533	0.956	0.694
21	Spain	0.942	-0.157	1.132	-0.618	0.891	-0.679	1.417	-0.533	0.956	0.694
22	Iceland	0.942	-0.157	1.132	-0.618	0.891	-0.679	1.417	-0.533	0.956	0.694
23	Brazil	0.942	-0.157	1.132	-0.618	0.915	-0.106	1.417	-0.533	0.956	0.694
24	Mexico	0.942	-0.157	0.629	0.004	0.915	-0.106	1.417	-0.533	0.956	0.694
25	Hungary	0.942	-0.157	0.629	0.004	0.891	-0.679	1.417	-0.533	0.956	0.694
26	Switzerland	0.942	-0.157	1.132	-0.618	0.915	-0.106	1.417	-0.533	0.956	0.694
27	Sweden	0.942	-0.157	0.905	-1.359	0.891	-0.679	1.417	-0.533	0.956	0.694
28	Denmark	0.942	-0.157	1.132	-0.618	0.891	-0.679	1.417	-0.533	0.956	0.694
29	Japan	0.942	-0.157	1.132	-0.618	0.891	-0.679	1.417	-0.533	0.956	0.694
30	Russian Federation	0.942	-0.157	0.629	0.004	0.891	-0.679	1.417	-0.533	0.956	0.694
31	Germany	0.942	-0.157	0.905	-1.359	0.891	-0.679	1.417	-0.533	0.956	0.694
32	New Zealand	0.942	-0.157	1.132	-0.618	0.915	-0.106	1.417	-0.533	0.956	0.694

Average P+ of Booklets 1, 2, & 6 by Country

