Plausible Values: How to Deal with Their Limitations

Christian Monseur
Australian Council for Educational Research

Raymond Adams
University of Melbourne

Rasch modeling and plausible values methodology were used to scale and report the results of the Organization for Economic Cooperation and Development's Programme for International Student Achievement (PISA).

This paper will describe the scaling approach adopted in PISA. In particular it will focus on the use of *plausible values,* a multiple imputation approach that is now commonly used in large-scale assessment. As with all imputation models the plausible values must be generated using models that are *consistent* with those used in subsequent data analysis. In the case of PISA the plausible value generation assumes a *flat* linear regression with all students' background variables collected through the international student questionnaire included as regressors. Further, like most linear models, homoscedasticity and normality of the conditional variance are assumed.

This paper will explore some of the implications of this approach. First, we will discuss the conditions under which the secondary analyses on variables not included in the model for generating the plausible values might be biased.

Secondly, as plausible values were not drawn from a multi-level model, the paper will explore the adequacy of the PISA procedures for estimating variance components when the data have a hierarchical structure.

Introduction

The Organization for Economic Cooperation and Development's Programme for International Student Achievement (PISA) is a survey of the Reading, Mathematics and Science proficiencies of 15-year-olds who are enrolled in school. PISA is an ongoing data collection that will assess students every three years. The first PISA data collection occurred in 2000 and the second occurred in 2003. The next data collection will occur in 2006. The international results of PISA 2000 and 2003 were published in December 2001 and 2004 respectively (OECD, 2001; OECD 2004).

To reach satisfactory coverage, many items need to be developed and included in the final test. At the same time, it is unreasonable and perhaps undesirable to assess a sampled student with the whole battery, therefore PISA implements a rotated test design. For details on the design see the initial report (OECD, 2001) and the study's technical report (Adams and Wu, 2002).

The purpose of a study such as PISA is to describe the characteristics of populations of the 15-year-olds students in school. That is, *the assignment of valid and reliable scores to individuals is not a purpose of PISA*. When the purpose of assessment is to describe populations rather than to measure individuals a number of authors (Mislevy, 1991; Adams, Wilson and Wu, 1992; Wu and Adams 2002), have shown that scaling the data with traditional item response methods, assigning students scale scores[1] and then analyzing the estimated scale scores to estimate population characteristics does not, in general, provide correct results. The alternative, and now generally preferred approach in studies such as PISA, The Third Intenational Mathematics and Science Study (TIMSS; Macaskill, Adams and Wu, 1998) and the National Assessment of Educational Progress (NAEP; Beaton, 1987), is the use of Mislevy's so-called plausible value methodology (Mislevy, 1991; Mislevy et al., 1992).

In this paper we describe the scaling methodology that was applied to PISA 2000. We then note and discuss two potential weaknesses in the methodology. These two weaknesses apply not only to the scaling of PISA, but also the scaling of TIMSS and NAEP. We then describe a set of simulations that were undertaken to examine the potential impact of these weaknesses on the outcomes of the PISA 2000 scaling.

PISA Scaling Methodology

The PISA data were scaled with the mixed coefficients multinomial logit model as described by Adams, Wilson and Wang (1997). The scaling was implemented using the ConQuest software (Wu, Adams and Wilson, 1997).

The model that was applied to the scaling of PISA was a generalized form of the Rasch model. The model is a mixed coefficients model because the items are described by a fixed set of unknown parameters ξ, while the student outcome levels (the latent variable), θ, is a random effect.

Assume that I items are indexed $i = 1,..., I$ with each item admitting $K_i + 1$ response categories indexed $k = 0,1,..., K_i$. Use the vector valued random variable, $\mathbf{X}_i = \left(X_{i1}, X_{i2},..., X_{iK_i}\right)^T$, where

$$X_{ij} = \begin{cases} 1 & \text{if response to item } i \text{ is in category } j \\ 0 & \text{otherwise} \end{cases}, (1)$$

to indicate the $K_i + 1$ possible responses to item i.

A vector of zeroes denotes a response in category zero. This effectively makes the zero category a reference category and is necessary for model identification. The choice of this as the reference category is arbitrary and does not affect the generality of the model. We can also collect the $\mathbf{X}_i$ together into the single vector $\mathbf{X}^T = \left(\mathbf{X}_1^T, \mathbf{X}_2^T,..., \mathbf{X}_I^T\right)$ which we call the response vector (or pattern). Particular instances of each of these random variables are indicated by their lower case equivalents; $\mathbf{x}$, $\mathbf{x}_i$ and x_{jk}.

The items are described through a vector $\xi^T = (\xi_1, \xi_2,..., \xi_p)$, of p parameters. Linear combinations of these are used in the response probability model to describe the empirical characteristics of the response categories of each item. Design vectors $\mathbf{a}_{ij}$, $(i = 1,...,I; j = 1,...,K_i)$, each of length $p,$ which can be collected to form a design matrix

---

$\mathbf{A}^T = \left(\mathbf{a}_{11}, \mathbf{a}_{12}, \ldots, \mathbf{a}_{1K_1}, \mathbf{a}_{21}, \ldots, \mathbf{a}_{2K_2}, \ldots, \mathbf{a}_{IK_I} \right)$ define these linear combinations.

The multidimensional form of the model assumes that a set of D traits underlie the individuals' responses. The D latent traits define a D-dimensional latent space and the vector, $\boldsymbol{\theta} = \left(\theta_1, \theta_2, \ldots, \theta_D \right)'$, represents the individuals' positions in the D-dimensional latent space.

An additional feature of the model is the introduction of a scoring function that allows the specification of the score or 'performance level' that is assigned to each possible response category to each item. To do this we introduce the notion of a response score b_{ijd} which gives the performance level of an observed response in category j of item i in dimension d. The scores across D dimensions can be collected into a column vector $\mathbf{b}_{ik} = \left(b_{ik1}, b_{ik2}, \ldots, b_{ikD} \right)^T$, and then again be collected into the scoring sub-matrix for item i, $\mathbf{B}_i = \left(\mathbf{b}_{i1}, \mathbf{b}_{i2}, \ldots, \mathbf{b}_{iD} \right)^T$, and then collected into a scoring matrix $\mathbf{B} = \left(\mathbf{B}_1^T, \mathbf{B}_2^T, \ldots, \mathbf{B}_I^T \right)^T$ for the whole test. (By definition, the score for a response in the zero category is zero, but other responses may also be scored zero).

The probability of a response in category j of item i is modelled as

$$\Pr\left(\mathbf{X}_{ij} = 1; \mathbf{A}, \mathbf{B}, \xi \mid \theta \right) = \frac{\exp\left(\mathbf{b}_{ij}\theta + \mathbf{a}'_{ij}\xi \right)}{\sum_{k=1}^{K_i} \exp\left(\mathbf{b}_{ik}\theta + \mathbf{a}'_{ik}\xi \right)} \quad (2)$$

And for a response vector we have;

$$f\left(\mathbf{x}; \xi \mid \theta \right) = \psi(\theta, \xi)\exp\left[\mathbf{x}'\left(\mathbf{B}\theta + \mathbf{A}\xi \right) \right], \quad (3)$$

with

$$\psi(\theta, \xi) = \left\{ \sum_{z \in \Omega} \exp\left[\mathbf{z}^T \left(\mathbf{B}\theta + \mathbf{A}\xi \right) \right] \right\}^{-1}, \quad (4)$$

where Ω is the set of all possible response vectors.

The Population Model

The item response model is a conditional model, in the sense that it describes the process of generating item responses conditional on the latent variable, θ. The complete definition of the model, therefore, requires the specification of a density, $f_\theta(\theta; \alpha)$ for the latent variable, θ. We use α to symbolise a set of parameters that characterise the distribution of θ. The most common practice when specifying uni-dimensional marginal item response models is to assume that the students have been sampled from a normal population with mean μ and variance σ^2. That is:

$$f_\theta\left(\boldsymbol{\theta}; \alpha \right) \equiv f_\theta\left(\theta; \mu, \sigma^2 \right) = \left(2\pi\sigma^2 \right)^{-\frac{1}{2}} \exp\left[-\frac{(\theta - \mu)^2}{2\sigma^2} \right], \quad (5)$$

or equivalently

$$\theta = \mu + E, \quad (6)$$

where $E \sim N(0, \sigma^2)$.

Adams, Wilson and Wu (1997) discuss how a natural extension of (5) is to replace the mean, μ with the regression model, $\mathbf{Y}_n^T \boldsymbol{\beta}$ where $\mathbf{Y}_n$ is a vector of u, fixed and known values for student n, and β is the corresponding vector of regression coefficients. For example, $\mathbf{Y}_n$ could be constituted of student variables such as gender or socioeconomic status. Then the population model for student n, becomes,

$$\theta_n = \mathbf{Y}_n^T \boldsymbol{\beta} + E_n, \quad (7)$$

where we assume that the E_n are independently and identically normally distributed with mean zero and variance σ^2 so that (7) is equivalent to:

$$f_\theta\left(\theta_n; \mathbf{Y}_n, b, \sigma^2 \right) = \left(2\pi\sigma^2 \right)^{-1/2}$$
$$\exp\left[-\frac{1}{2\sigma^2}\left(\theta_n - \mathbf{Y}_n^T \boldsymbol{\beta} \right)^T \left(\theta_n - \mathbf{Y}_n^T \boldsymbol{\beta} \right) \right], \quad (8)$$

a normal distribution with mean $\mathbf{Y}_n^T \boldsymbol{\beta}$ and variance σ^2. If (8) is used as the population model then the parameters to be estimated are β, σ^2 and ξ.

The generalization needs to be taken one step further to apply it to the vector valued, θ, rather than the scalar valued θ. The extension results in the multivariate population model:

$$f_\theta\left(\boldsymbol{\theta}_n; \mathbf{W}_n, \boldsymbol{\gamma}, \boldsymbol{\Sigma} \right) = \left(2\pi \right)^{-d/2} \left| \boldsymbol{\Sigma} \right|^{-\frac{1}{2}}$$
$$\exp\left[-\frac{1}{2}\left(\boldsymbol{\theta}_n - \boldsymbol{\gamma}\mathbf{W}_n \right)^T \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\theta}_n - \boldsymbol{\gamma}\mathbf{W}_n \right) \right], \quad (9)$$

where γ is a $u \times d$ matrix of regression coefficients, Σ is a $d \times d$ variance-covariance matrix and $\mathbf{W}_n$ is a $u \times 1$ vector of fixed variables. In PISA the $\mathbf{W}_n$ variables as referred to as conditioning variables.

Combined Model

In (10), the conditional item response model (6) and the population model (9) are combined to obtain the unconditional, or marginal, item response model

$$f_x(\mathbf{x};\boldsymbol{\xi},\gamma,\Sigma) = \int_\theta f_x(\mathbf{x};\boldsymbol{\xi}\,|\,\theta)\, f_\theta(\theta;\gamma,\Sigma)\, d\theta. \quad (10)$$

It is important to recognize that under this model the locations of individuals on the latent variables are not estimated. The parameters of the model are γ, Σ and ξ.

The procedures that are used to estimate the parameters of the model are described in Adams, Wilson and Wu (1997), and Adams, Wilson, and Wang (1997).

For each individual it is possible however to specify a posterior distribution for the latent variable. The posterior distribution is given by:

$$\begin{aligned} h_\theta(\theta_n;\mathbf{W}_n,\boldsymbol{\xi},\gamma,\Sigma\,|\,\mathbf{x}_n) &= \frac{f_x(\mathbf{x}_n;\boldsymbol{\xi}\,|\,\theta_n)\, f_\theta(\theta_n;\mathbf{W}_n,\gamma,\Sigma)}{f_x(\mathbf{x}_n;\mathbf{W}_n,\boldsymbol{\xi},\gamma,\Sigma)} \\ &= \frac{f_x(\mathbf{x}_n;\boldsymbol{\xi}\,|\,\theta_n)\, f_\theta(\theta_n;\mathbf{W}_n,\gamma,\Sigma)}{\int_{\theta_n} f_x(\mathbf{x}_n;\boldsymbol{\xi}\,|\,\theta_n)\, f_\theta(\theta_n;\mathbf{W}_n,\gamma,\Sigma)} \end{aligned} \quad (11)$$

Application To PISA

In PISA this model has been used in three steps:

- national calibrations;
- international scaling; and
- student score generation.

For both the national calibrations and the international scaling the conditional item response model (3) is used in conjunction with the population model (9), but conditioning variables are not used. That is, it is assumed that students have been sampled from a multivariate normal distribution.

For PISA 2000 the model was a five-dimensional model, made up of three reading dimensions, one science and one mathematics dimension. The design matrix was chosen so that the partial credit model was used for items with multiple score categories and the simple logistic model was fitted to the dichotomously scores items.

National Calibrations

The national calibrations were performed separately country-by-country using unweighted data. The results of these analyses were used to monitor the quality of the data and to make decisions regarding national item treatment.

- an item was deleted from PISA if it had poor psychometric characteristics in more than 8 countries.

- an item would be regarded as not-administered in particular countries if it had poor psychometric characteristics in that country, but functions well in the vast majority of others.

- an item that has sound characteristics in each country but shows substantial item-by-country interactions may be regarded as a different (for scaling purposes) item in each country (or in some subset of the countries). That is, the difficulty parameter will be free to vary across countries.[2]

Note that both the second and third options above have the same impact on comparisons between countries. That is, if an item is identified as behaving differently in one country than in others, then choosing either the second or third of the above options will have the same impact on inter-country comparisons. The choice between the second or third options could, however, influence within-country comparisons.

When reviewing the national calibrations particular attention was paid to the fit of the items to the scaling model, item discrimination and item-by-country interactions.

International Calibration

International item parameters were set by applying the conditional item response model (3) in conjunction with the multivariate population model (9), without the use of conditioning

---

2   No decisions of this type were taken in PISA 2000.

variables to a sub-sample of students. This sub-sample of students, referred to as the international calibration sample, consisted of 500 students drawn at random from each of the participating OECD countries.

Student Score Generation

As with all item response scaling models the *proficiencies* (or measures) of each student are not observed. The proficiencies are missing data that must be inferred from the observed item responses.

There are a number of possible alternative approaches for inferring these missing proficiencies. In PISA, we have used two approaches: maximum likelihood, using Warm's (1985) weighted estimator (WLEs), and plausible values (PVs).

• The WLE proficiency is the proficiency that makes the score that the student attained most likely.

• The PVs are a selection of likely proficiencies for students that attained each score.

Computing Maximum Likelihood Estimates in PISA

In PISA 2000, six weighted likelihood estimates were provided for each student, one for each of mathematics literacy, reading literacy and scientific literacy and one for each of three reading literacy sub-scales.[3]

Weighted maximum likelihood ability estimates (WLE; Warm, 1985) are produced by maximising (3) with respect to θ_n, that is, solving the likelihood equations

$$\sum_{d \in D} \left[\sum_{i \in \Omega} \left(\mathbf{b}_{i x_{ni}} - \sum_{j=1}^{K_i} \frac{\mathbf{b}_{ij} \exp\left(\mathbf{b}_{ij} s_{nd} \theta_{nd} + \mathbf{a}'_{ij} \hat{\boldsymbol{\xi}}\right)}{\sum_{k=1}^{K_i} \exp\left(\mathbf{b}_{ik} s_{nd} \theta_{nd} + \mathbf{a}'_{ik} \hat{\boldsymbol{\xi}}\right)} \right) + \frac{J_{nd}}{2 I_{nd}} \right] = 0, \ (12)$$

for each case, where $\hat{\boldsymbol{\xi}}$ are the item parameter estimates obtained from the international calibration and d indicates the latent dimensions. I_{nd} is the test information for student n on dimension d and J_{nd} the first derivative with respect to theta.

These equations are solved using a routine based on the Newton-Raphson method.

Plausible Values

Using item parameters anchored at their estimated values from the international calibration plausible values are random draws from the marginal posterior of the latent distribution, (9), for each student. For details on the uses of plausible values the reader is referred to Mislevy (1991) and Mislevy et al. (1992).

For PISA the random draws from the marginal posterior distribution, (11), are taken as follows.

M vector-valued random deviates, $\left\{ \boldsymbol{\varphi}_{mn} \right\}_{m=1}^{M}$, from the multivariate normal distribution, $f_{\theta}\left(\boldsymbol{\theta}_n; \mathbf{W}_n, \boldsymbol{\gamma}, \Sigma\right)$, for each case n.[4] These vectors are used to approximate the integral in the denominator of (11), using the Monte-Carlo integration

$$\int_{\theta} f_{\mathbf{x}}\left(\mathbf{x}; \boldsymbol{\xi} \mid \boldsymbol{\theta}\right) f_{\theta}\left(\boldsymbol{\theta}, \boldsymbol{\gamma}, \Sigma\right) d\boldsymbol{\theta}$$

$$\approx \frac{1}{M} \sum_{m=1}^{M} f_{\mathbf{x}}\left(\mathbf{x}; \boldsymbol{\xi} \mid \boldsymbol{\varphi}_{mn}\right) \equiv \Im. \quad (13)$$

At the same time, the values

$$p_{mn} = f_{\mathbf{x}}\left(\mathbf{x}_n; \boldsymbol{\xi} \mid \boldsymbol{\varphi}_{mn}\right) f_{\theta}\left(\boldsymbol{\varphi}_{mn}; \mathbf{W}_n, \boldsymbol{\gamma}, \Sigma\right) (14)$$

are calculated, so that we obtain the set of pairs $\left(\boldsymbol{\varphi}_{mn}, p_{mn}/\Im\right)_{m=1}^{M}$, which can be used as an approximation of the posterior density (11); and the probability that φ_{nj} could be drawn from this density is given by

$$q_{nj} = \frac{p_{mn}}{\sum_{m=1}^{M} p_{mn}}. \quad (15)$$

At this point, L uniformly distributed random numbers, $\left\{ \eta_i \right\}_{i=1}^{L}$, are generated; and for each random draw, the vector, φ_{nj}, that satisfies the condition

---

3   Note that in PISA 2003 weighted likelihood estimates were not provided since, unlike plausible values, they could not be estimated for all sampled students.

4   The value M should be large. For PISA we have used 2000.

$$\sum_{s=1}^{i_0-1} q_{sn} < \eta_i \leq \sum_{s=1}^{i_0} q_{sn} \qquad (16)$$

is selected as a plausible vector.

It is important to recognize that this approach to drawing the plausible values differs slightly from that implemented in NAEP. The NAEP procedures use an approximate method that is based on assuming that the posterior distribution (11) is normal (Beaton, 1987; Thomas and Gan, 1997).[5]

Constructing Conditioning Variables

The PISA conditioning variables are prepared using procedures based upon those used in the NAEP (Beaton, 1987) and the TIMSS (Macaskill, Adams and Wu, 1998).

The steps involved in this process are as follows:

Step 1: Each variable in the student question-naire was dummy coded. This is undertaken so that categorical variables can be used in the following steps and so that missing data can be included.

Step 2: For each country a principal components analysis of the dummy coded variables was performed and component scores were produced for each student. A sufficient number of components to account for 90 per cent of the variance in the original variables were retained.

Step 3: Using item parameters anchored at their international location and conditioning variables derived from the national principal components analysis the item response model was fitted to each national data set and the national population parameters γ, and Σ were estimated.[6]

Step 4: Five vectors of plausible values are drawn using the method described above.

---

[5]   Chang and Stout (1993) show that the posterior is asymptotically normal. But as the number of items taken by individual students is often very small we do not believe it is an appropriate assumption to make in PISA.

[6]   In addition to the principal components gender, ISEI and school mean performance were added as conditioning variables.

The Analysis of Data with Plausible Values

It is very important to recognise that plausible values are *not* test scores and should not be treated as such. Plausible values are random numbers that are drawn from the distribution of scores that could be reasonably assigned to each individual—that is from the marginal posterior distribution (11). As such the plausible values contain random error variance components and are not optimal as scores for individuals. The beauty of plausible values is that as a set they are better suited to describing the performance of the population than is a set of scores that are optimal at the individual student level.[7]

The plausible value approach, which was developed by Mislevy and Sheehan (1989) based upon the imputation theory of Rubin (1987), produces consistent estimators of population parameters provided that the imputation model is commensurate with the data analytic model.

Plausible values are intermediate values that are provided so that consistent estimates of population parameters can be obtained using standard statistical analysis software such as SPSS and SAS. As an alternative to plausible values, analyses can be completed using a package such as ConQuest (Wu, Adams and Wilson, 1997).

PISA provides five plausible values per scale or subscale. If an analysis is to be undertaken with one of these five cognitive scales then (ideally) the analysis should be undertaken five times, once with each of the five relevant plausible values variables. The results of these five analyses are averaged and then significance tests that adjust for variation between the five sets of results are computed.

More formally, suppose that $r(\theta,\mathbf{Y})$ is some statistic that depends upon the latent variable and some other observed characteristic of each student. That is: $(\theta,\mathbf{Y}) = (\theta_1, y_1, \theta_2, y_2, \ldots, \theta_N, y_N)$ where θ_N, y_N are the values of the latent variable and the other observed characteristic for student n. Unfortunately θ_n is not observed, although we do

---

[7]   Where optimal might be defined for example as either unbiased or minimising the mean squared error at the student level.

observe the item responses, x_n from which we can construct for each student, n, the marginal posterior $h_\theta(\theta_n; y_n, \xi, \gamma, \Sigma | x_n)$. So that if $h_\theta(\theta; Y, \xi, \gamma, \Sigma | X)$ is the joint marginal posterior for n = 1,...,N then we can compute:

$$r^*(\mathbf{X}, \mathbf{Y}) = E\left[r^*(\theta, \mathbf{Y}) | \mathbf{X}, \mathbf{Y}\right]$$
$$= \int_\theta r(\theta, \mathbf{Y}) h_\theta(\theta; \mathbf{Y}, \xi, \gamma, \Sigma | \mathbf{X}) d\theta. \quad (17)$$

The computation of the integral in (17) can be accomplished using the Monte-Carlo method. If M random vectors $(\Theta_1, \Theta_2, ..., \Theta_M)$ are drawn from $h_\theta(\theta; Y, \xi, \gamma, \Sigma | X)$ (17) is approximated by:

$$r^*(\mathbf{X}, \mathbf{Y}) \approx \frac{1}{M} \sum_{m=1}^M r(\Theta_m, \mathbf{Y})$$
$$= \frac{1}{M} \sum_{m=1}^M \hat{r}_m, \quad (18)$$

where $\hat{r}_m$ is the estimate of r computed using the mth set of plausible values.

From (17) we can see that the final estimate of r is the average of the estimates computed using each plausible value in turn. If U_m is the sampling variance for $\hat{r}_m$ then the sampling variance of r^* is:

$$V = U^* + \left(1 + M^{-1}\right) B_M, \quad (19)$$

where

$$U^* = \frac{1}{M} \sum_{m=1}^M U_m$$

and

$$B_M = \frac{1}{M-1} \sum_{m=1}^M \left(\hat{r}_m - r^*\right)^2.$$

An α-% confidence interval for r^* is

$$r^* \pm t_v\left(\frac{(1-\alpha)}{2}\right) V^{\frac{1}{2}}$$

where $t_v(s)$ is the s percentile of the t-distribution with n degrees of freedom.

$$v = \frac{1}{\dfrac{f_M^2}{M-1} + \dfrac{(1-f_M)^2}{d}},$$
$$f_M = \left(1 + M^{-1}\right) B_M / V$$

and d is the degrees of freedom that would have applied if θ_n had been observed. In PISA the value of d, will vary from country-to-country and will have a maximum possible value of 80.

Note that analyses based upon a single plausible value will provide unbiased results (if the imputation model is correct), but will underestimate the error variance since the measurement error component cannot be included.

Possible Issues with the Methodology

It was noted above that a plausible value approach would provide consistent estimates of population parameters if the imputation model that is used to generate the plausible values is commensurate with the data analytic model that-There are many ways there may be an inconsistency between the imputation and data analytic models, but there are few published studies that explore this issue. One example is Thomas (2000), which examines the influence of skewness in the latent distribution on the estimation of means, variances and percentiles.

In this paper we explore two ways in which the imputations and data analytic models may differ. First, the sampling model for PISA, and for most school-based studies, typically involves the use of schools as the primary sampling unit (selected with probability proportional to size) followed by random sampling of a fixed number of eligible students from each sampled school. Because the students are nested within schools, and in most education systems are not randomly assigned to schools, a two-level model will often be the appropriate method for data analysis. The population model underlying the plausible value generation (equation (8)) is a single-level model that does not recognise the hierarchical structure of the data.

Second, the *conditioning* variables used do not represent all possible variables that might be used as regressors in subsequent analysis. It does not, for example, include interactions between independent variables, nor does it include national variables that countries may add to the instrumentation.

In following simulations we report the extent to which PISA results might be compromised by the failure of the conditioning model to deal with these two issues.

Simulations: One

The first set of simulations intends to analyse the effectiveness of different proficiency estimators for use in recovering the decomposition of the sampling variance in the case of a cluster sampling design like that used in PISA. We compare four approaches. They are the usual maximum likelihood estimator (MLE), the WLE (or weighted likelihood estimator) described above, the expected a-posteriori estimator (EAP) and plausible values (PVs) (as described above).

Wu and Adams (2002) showed that estimates of the population mean computed using MLEs, WLEs, EAPs and PVs are not biased, but that estimates of population variance are underestimated if the EAP is used, overestimated if the MLE or WLE is used and correctly estimated with PVs.

In this first simulation we explore the effectiveness of various estimators of the within- and between-school variance in a two stage, PISA-like design. The decomposition of variance into these components is important for at least two reasons. First, the relationship between these two variance components is often regarded as an important descriptive statistic. Second, these components may be needed to estimate the sampling variance of other statistics, such as the sampling mean.

In a two stage cluster design where a simple and random sample of schools is drawn from an infinite population of schools and equal size simple and random samples of students is drawn from infinite populations of students within each school, the sampling variance on the mean is equal to:

$$\sigma_{(\hat{\mu})}^2 = \frac{\sigma_b^2}{n_b} + \frac{\sigma_w^2}{n_b \times n_w}, \qquad (20)$$

where σ_b^2 is the between cluster variance, (that is, the school variance), σ_w^2 is the within-school variance, n_b is the number of schools in the sample and n_w the number of students per school. As

this formula clearly shows, biased estimates of the cluster and within cluster variances will lead to biased estimates of the sampling variance on the mean.

Design of the simulation

As the reliability of the test might affect the estimation of the cluster and within cluster variance estimates and as the reliability is partly a function of the number of items in the test, the simulations were performed with different number of dichotomous items, respectively 3, 20, 50 and 100.

The simulations were designed so that Con-Quest (Wu, Adams and Wilson, 1997) generates for each simulation and for each replicate a sample of 80 schools and for each school, a sample of 25 students. This gives a total of 2000 students.

For data generation the between-school variance was set at 0.4 and the within-school variance was set at 0.6 and item parameters were randomly generated from a uniform distribution with minimum –2.0 and maximum 2.0. After generation the mean of the item parameters was set at zero.

The number of replicates for each of the four simulations was chosen to ensure a standard error of 0.005 on the estimator of the population mean based upon plausible values. The number of replicates was set at 276 for a test of three items, to 223 for a test of 20 items, to 219 for a test of 50 items and finally to 213[8] for a test of 100 items. The difference in the number of replicates reflects the variation in the measurement error depending on the number of items included in the tests.

For each replicate the item and person parameters were generated as above and then item response vectors that conformed to the dichotomous Rasch model were generated. The marginal Rasch model (10) was then estimated from the data and MLE, WLE, EAP and PVs were computed for each case.[9]

---

8   Due to time constraints, the simulation with conditioning for a test of 100 items is based on 150 replications.

9   Both data generation and estimation were undertaken with a new version on of the ConQuest software, which provides a wide variety of data generation and simulation features.

Note that, in this first set of four simulations conditioning variables were not used. To emulate the PISA procedures a second set of simulations was then run with school mean of the student latent proficiencies used as a conditioning variable.

For each simulation, the mean and the variance of the different type of estimates were computed. The between-school variance and the within-school variance were also computed for each replicate using SAS PROC MIXED (Littell, Milliken, Stroup, Wolfinger, 1999).

Most of our results are presented in terms of the means of the between-school and within-school variance estimates across replications. The variation across replicates of the between-school and within-school variances is used to estimate their respective sampling distribution errors.

We have used z-tests to compare the difference of mean of the estimates with the generating values—namely, 0.4 for the between-school variance and 0.6 for the within-school variance.[10]

Results

Table 1 presents the means of the between-school variance estimates for each type of proficiency estimates and for each of the eight simulations and tests of their differences from the generating values. The first five rows show results for the simulations without conditioning and the following three rows show the results with conditioning. The MLE and WLE results are not affected by the choice of conditioning variables.

The MLEs underestimate the between-school variance with a test of 3 items but overestimate the school variance once the test contains a larger number of items. The WLEs underestimate the between-school variance with a test of 3 items but appear to do a reasonable job when there are 20 items or more.

It should be noted that these conclusions are only valid for the conditions of these simulations—that is, an intraclass correlation of 0.4 and clusters of 25. While it is not expected that the number of schools in the sample would directly affect the bias of the estimates,[11] it is expected that a reduction of the within school sample size or a modification of the intraclass correlation would influence the bias. The bias should be inversely proportional to the number of students selected per school.

The EAP and Plausible values, whether the estimates are based on one plausible value (PV1) or the mean estimates of 5 Plausible Values (5PV) clearly underestimate the school variances. The bias appears to be proportional to the number of items in the test, in other words, to the reliability of the test. EAP and PVs behave in a similar way. The school estimates are on average identical.

The EAP and PVs school variance estimates that were generated by conditioning on the latent proficiency estimates are all unbiased, even with a test of three items.

Table 2 presents the means of the within-school variance estimates for each type of proficiency estimate and for each of the eight

---

10   The large degrees of freedom for the variance estimates means they should be very closely to normally distributed under the null hypothesis.

11   Unless the school sample is be very small.

Table 1

Means of the Between-School Variance Estimates

	3 items		20 items		50 items		100 items	
	Mean	Z	Mean	z	Mean	z	Mean	z
MLE	0.327	−19.50	0.445	8.75	0.415	3.14	0.407	1.36
WLE	0.272	−38.08	0.399	−0.16	0.399	−0.25	0.399	−0.26
EAP	0.050	−305.74	0.241	−50.06	0.320	−20.19	0.356	−10.51
PV1	0.049	−270.72	0.241	−48.37	0.319	−20.43	0.356	−10.30
5 PV	0.050	−242.12	0.241	−47.24	0.320	−19.95	0.356	−9.99
EAP With Conditioning	0.392	−1.57	0.399	−0.38	0.393	−1.47	0.397	−0.47
PV1 With Conditioning	0.393	−1.50	0.397	−0.76	0.392	−1.49	0.397	−0.53
5PV With Conditioning	0.392	−1.44	0.398	−0.51	0.392	−1.48	0.397	−0.49

simulations and tests of their differences from the generating values. As Table 2 shows, most of the within school variance estimates are substantially biased. The upward bias in the MLE and WLE based estimates of the total variance (Wu and Adams, 2001) mostly affect the within-school variance estimates. As was observed for the total variance the bias is a function of the reliability of the test. For instance, the mean of the WLE between-school variance estimates is equal to 0.75 for a test of 50 items. The EAP, with or without conditioning, systematically underestimate the within school variance while the EAP with conditioning recovers the school variance. Finally, plausible values without conditioning overestimate the within-school variance. This result is not unexpected, as the plausible value methodology provides unbiased estimates of the total variance and as the between-school variance is substantially underestimated the within-school variance must be overestimated. The best estimates of the within-school variance are based on plausible values generated with conditioning

on the latent proficiency school means. In this case the normalized difference with the expected values seems to indicate a slight overestimation.

Table 3 summarises the estimation of the relationship between the two variance components. It does so by reporting the intraclass correlation coefficients. As the generating between-school variance was 0.4 and the generating within-school variance was 0.6, the generating intraclass correlation is equal to 0.4.

MLEs, WLEs, EAPs and PVs generated without conditioning systematically underestimate the intraclass correlation. The bias appears to be directly related to the reliability of the test, but as noted previously, other parameters like the expected intraclass correlation or the within school sample size should also influence the bias. Nevertheless, these two other parameters should not change the relationship between the reliability of the test and the size of the bias.

In the introduction of the first set of simulations, the importance of reliably estimating the

Table 2

Means of the Within-School Variance Estimates

	3 items		20 items		50 items		100 items	
	Mean	Z	Mean	z	Mean	z	Mean	z
MLE	2.037	130.38	1.010	151.68	0.753	82.57	0.672	38.74
WLE	1.711	71.83	0.901	125.19	0.722	69.78	0.659	32.43
EAP	0.302	−71.63	0.537	−30.57	0.576	−14.35	0.588	−6.70
PV1	0.947	43.16	0.762	58.85	0.683	42.84	0.643	21.91
5 PV	0.948	42.04	0.763	54.45	0.682	38.16	0.643	20.90
EAP With Conditioning	0.151	−161.99	0.407	−106.39	0.505	−61.91	0.555	−22.99
PV1 With Conditioning	0.615	2.18	0.606	2.52	0.604	2.19	0.608	4.13
5PV With Conditioning	0.612	1.85	0.606	2.46	0.604	2.00	0.609	4.47

Table 3

Intraclass Correlation Estimates

	3 items	20 items	50 items	100 items
MLE	0.138	0.306	0.355	0.377
WLE	0.137	0.307	0.356	0.377
EAP	0.141	0.310	0.357	0.377
PV1	0.049	0.240	0.318	0.356
5 PV	0.050	0.240	0.319	0.356
EAP With Conditioning	0.722	0.495	0.437	0.417
PV1 With Conditioning	0.390	0.396	0.394	0.395
5PV With Conditioning	0.391	0.396	0.394	0.395

between-school and within-school variance for estimating the sampling variance on the mean for instance, was mentioned. The results reported here show that plausible values can provide unbiased estimates of the school and within-school variances if conditioning on the school mean is implemented. Neither the MLEs, WLEs, nor EAPs can be used for this purpose. If the between-school or the within-school variances are biased, one might therefore expect biases in the sampling variance on the mean estimates. To express the importance of the bias, the ratio of the standard error estimate on the mean estimate to the standard error was computed.[12] This ratio was then multiplied by 1.96 and the α-error computed. If the sampling variance estimate is unbiased, α will be equal to 0.05. If the sampling variance on the mean is underestimated, α will be greater than 0.05. This means that the null hypothesis will be more often rejected than it should be. Values lower than 0.05 mean that the null hypothesis will be less often rejected than expected. One might be pleased with this latter outcome but on the other hand, it increases the Type 2 error, that is accepting the null hypothesis when it is false.

Table 4 presents the Type 1 error per simulation per type of proficiency estimate. The biggest biases are associated with EAP and PV generated without conditioning on the school mean of the latent proficiency. As the sampling variance on a

---

12   Strictly speaking one is interested in the ratio of the estimate of the combined sampling and measurement error variance to the between replication variance. This may mean that the following results are based on ratios that are very slight underestimated for the PVs.

mean for a cluster sample mainly depends on the cluster variance, these results reflect the important underestimation of the cluster variance and it demonstrates the importance of the conditioning. As the EAPs generated with conditioning properly estimate the between-school variance but systematically underestimate the within-school variance, the Type 1 error is only slightly overestimated. PVs with conditioning seem the most appropriate, as this type of estimates can recover both variance in a satisfactory way. Finally, while MLEs and WLEs proved to provide poor estimates of the variances, and one might have expected large inconsistencies in Table 4 for these estimates, clearly this is not the case. As the bias is mostly located at the within-school level, type 1 error for MLE and especially for WLE are quite acceptable.

Simulations: Two

A national or an international assessment, such as PISA, does not only focus on the estimation of the proficiency mean for a population or a set of populations. Policy makers may use the outcomes of the assessment to initiate educational reforms supposed to improve the quality and the equity of the educational system.

Therefore, student proficiency estimates are usually related to some student background or behavioral characteristics or with some teacher or school features. In this context, it is important to ensure that the estimates based upon student proficiency estimates are good estimates of the relationship that exists between the latent proficiency and the school or student characteristics.

Table 4

Type 1 Error Rates

	3 items	20 items	50 items	100 items
MLE	0.054	0.036	0.045	0.048
WLE	0.079	0.047	0.049	0.050
EAP	0.454	0.123	0.078	0.064
PV1	0.374	0.117	0.077	0.063
5 PV	0.373	0.117	0.076	0.063
EAP With Conditioning	0.057	0.053	0.053	0.051
PV1 With Conditioning	0.052	0.051	0.052	0.051
5PV With Conditioning	0.052	0.050	0.052	0.051

As student proficiency estimates are contaminated by measurement error, it is well known that the first order correlation between student proficiency estimates and any school or student characteristics is biased towards zero (eg Guilford, 1954; Fuller, 1987).

Suppose Y is a variable measured without error and θ is a second variable measured with error by $\hat{\theta}$ then if $\hat{\theta} = \theta + \varepsilon$ and $\text{cov}\left(\hat{\theta}, \varepsilon\right) = 0$ then it is easy to show that an unbiased estimate of the correlation between Y and θ is given by:

$$r_{\theta Y} = \frac{r_{\hat{\theta} Y}}{\sqrt{r_{\hat{\theta}\hat{\theta}}}} \tag{21}$$

where $r_{\hat{\theta} Y}$ is the observed correlation between Y and $\hat{\theta}$, $r_{\hat{\theta}\hat{\theta}}$ is the reliability of $\hat{\theta}$, that is

$$r_{\hat{\theta}\hat{\theta}} = \frac{\sigma_\theta^2}{\sigma_{\hat{\theta}}^2}.$$

While these conditions hold for the WLE and MLE they do not hold for the EAP. In the case of the EAP, the estimated value and the error, ε, are correlated. However Wu and Adams (2002) show that for the EAP a result analogous to (21) holds if $r_{\hat{\theta}\hat{\theta}}$ is given by

$$r_{\hat{\theta}\hat{\theta}} = \frac{\sigma_{\hat{\theta}}^2}{\sigma_\theta^2}.$$

The second simulation intends to examine how well the different types of student proficiency estimates recover the latent correlation between the student proficiency and another characteristic.

As Guilford pointed out, the bias in the correlation estimate will depend on the reliability of the instrument. Therefore, as above, simulations were based on 3, 20, 50 and 100 items.

Simple random samples of 2000 students were drawn from a normal latent population distribution with a mean of zero and a standard deviation of one. The number of replicates used for the first simulation was used for the second one. So the results presented below are based on 276 replicates for a test of 3 items, 223 for a test

of 20 items, 220 for a test of 50 items are 213[13] for a test of 100 items.

Results

The observed correlations and correlation estimates corrected for attenuation are provided for these simulations in Table 5. The reliability of the MLE and WLE used for the dissatenuation are defined as:

$$r_{\hat{\theta}\hat{\theta}} = \frac{1}{\sigma_{\hat{\theta}}^2}. \tag{22}$$

For the EAP the dissattentuation was computed using:

$$r_{\hat{\theta}\hat{\theta}} = \sigma_{\hat{\theta}}^2. \tag{23}$$

The reliability of the plausible values was estimated as the mean of the 10 correlations between the five PVs. The average correlation was computed by using the Fisher transformation. Table 5 presents the average of the correlation coefficient estimates[14] with their standardized difference between the correlation and the expected value of 0.30.[15]

As expected, MLE, WLE, EAP and plausible values without conditioning provide biased estimates of the latent correlation. The bias is also proportional to the reliability of the test.

The correction for attenuation appears to work less well for the MLE than for WLE. Indeed, with a test of 20 items or higher, the correlation estimate corrected for attenuation is not significantly different from the generating value while more than fifty items seems necessary to provide an unbiased estimate of the latent correlation with MLE. This is likely due to the outward bias in MLEs.

---

[13]   Due to time constraints, the simulation for a test of 100 items is based on 202 replications with no conditioning, 135 replications with conditioning.

[14]   The Fisher transformation was firstly applied to the correlation estimates. The mean of the transformed correlations was computed and then the inverse transformation was applied.

[15]   The Fisher transformation was also used to test the difference with the expected value.

The correction for attenuation is best for EAP estimated without conditioning. Even with a test of 3 items, the correlation estimate is unbiased. On the other hand, the correction for attenuation does not work for plausible values generated without conditioning.

On average, the MLE, WLE and EAP correlation estimates appear to be identical. This result is not surprising, as in the case of a complete test design, MLE, WLE and EAP each associate a common estimate (in logits) to each possible raw score. Nevertheless, as the interval between two possible logit scores from one type of estimate to another one does not depend on a linear relationship, slight differences will be observed between the correlation estimates.

Plausible values with conditioning provide unbiased estimates of the latent correlation, regardless the number of items in the test. On the other hand, due to the underestimation of the total variance, and more specifically the underestimation of the residual variance, EAP estimates systematically overestimate the latent correlation. As the bias of the residual variance is a function of the reliability, then the product of the observed correlation by the reliability appears to be a satisfactory way of correcting this overestimation. We have not however yet proved this result. In our example, this would give correlation estimates of

0.280, 0.297 and 0.299 respectively for a test of 3, 20 and 50 items.

Discussion and Conclusion

In the PISA 2000 international database, two types of proficiency estimates were included: Weighted Maximum Likelihood Estimates and Plausible Values. Each student who took part in the assessment has at least results on the combined reading scale and the three reading subscales (retrieving information, interpreting and reflecting) and about 5/9 of the assessed students students also have proficiency estimates in mathematics and/or in science.

With an average test length of 60 items, it follows that student proficiency estimates for the reading subscales, the mathematics scale and the science scale are based on a small set of items (typically between 15 and 20).

As the database is currently widely used for secondary analysis, it is important to ensure that the type of proficiency estimates returned minimises the risk of bias.

PISA 2000 data were collected through a complex sampling design. In most cases, schools were firstly drawn and then students were randomly selected. Goldstein (1987), Bryk and Raudenbush (1992) brought to our attention the importance of recognising the hierarchical struc-

Table 5

Means of Correlation estimates

	3 items		20 items		50 items		100 items	
	Mean	Z	Mean	z	Mean	z	Mean	z
MLE	0.176	−99.88	0.262	−30.75	0.283	−14.40	0.291	−6.59
WLE	0.175	−100.87	0.262	−30.28	0.283	−14.25	0.292	−6.53
EAP	0.176	−99.91	0.263	−29.47	0.283	−14.09	0.292	−6.43
PV1	0.105	−160.76	0.230	−52.00	0.268	−26.97	0.283	−13.06
5 PV	0.105	−146.51	0.231	−51.95	0.268	−26.58	0.283	−12.61
MLE corrected for attenuation	0.272	−14.15	0.316	9.30	0.306	4.46	0.304	2.35
WLE corrected for attenuation	0.248	−26.81	0.299	−0.66	0.300	0.16	0.301	0.49
EAP corrected for attenuation	0.299	−0.49	0.299	−0.88	0.299	−0.69	0.300	−0.28
PV1 corrected for attenuation	0.178	−62.70	0.261	−25.73	0.283	−13.43	0.291	−6.51
EAP With Conditioning	0.478	47.68	0.339	24.61	0.317	12.88	0.309	5.86
PV1 With Conditioning	0.298	−1.01	0.298	−1.43	0.301	0.40	0.300	0.32
5PV With Conditioning	0.297	−1.18	0.299	−0.51	0.300	0.26	0.300	0.15

ture of the data. The first simulation has shown that the hierarchical structure of the data needs to be taken into account not only when secondary analyses are performed but also when student proficiency estimates are generated. The results show the superiority of the plausible values with conditioning for recovering the between-school and within-school variances.

The second simulation compares the efficiency of the different types of estimates for recovering a latent correlation. It appears that WLE will provide unbiased estimates if the test has at least about 20 items. But this correlation needs to be attenuated and users of the database might not be aware of this requirement. Plausible values allow the users to safely avoid this correction. Nevertheless, the generation of plausible values requires a careful conditioning; otherwise their superiority will be limited to the estimation of the total variance and percentiles.

Acknowledgement

References

Adams, R. J., and Wilson, M. R., and Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21, 1-24

Adams, R. J., and Wu, M. L. (Eds.) (2002) PISA 2000 Technical Report. OECD, Paris.

Adams, R. J., Wilson, M. R., and Wu, M. L. (1997) Multilevel item response modelling: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22,* 47-76.

Beaton, A. E. (1987). *Implementing the new design: The NAEP 1983-84 Technical Report.* (Report No. 15-TR-20). Princeton, NJ: Educational Testing Service.

Bryk, A. S., and Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park: Sage.

Chang, H., and Stout, W. (1993). The asymptotic posterior normality of the latent train in an IRT model. *Psychometrika*, *58*, 37-52.

Goldstein, H. (1987). *Multilevel statistical models*. London: Edward Arnold.

Guilford, J. P. (1954) *Psychometric methods*. New York: McGraw-Hill.

Fuller, W. (1987). *Measurement error models*. New York: Wiley.

Littell, R. C., Milliken, G. A., Stroup, W. W., and Wolfinger, R. D. (1999). *SAS system for mixed models*. Cary, NC: SAS Institute.

Macaskill, G., Adams, R. J., and Wu, M. L. (1998). Scaling methodology and procedures for the mathematics and science literacy, advanced mathematics and physics scales. In M. Martin and D. L. Kelly (Eds.) *Third international mathematics and science study. Technical Report Volume 3: Implementation and Analysis*. Boston College: Chestnut Hill, MA.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177-196.

Mislevy, R. J., Beaton, A. E., Kaplan, B., and Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*, 133-161.

Mislevy, R. J., and Sheehan, K. M. (1989). Information matrices in latent-variable models. *Journal of Educational Statistics*, 14(4), 335-350.

Organization for Economic Co-operation and Development. (2001). *Knowledge and skills for life: First results from PISA 2000*. Paris: OECD Publications.

Organization for Economic Co-operation and Development. (2004). *Learning for tomorrow's world: First results from PISA 2003*. Paris: OECD Publications.

Rubin, D. B. (1987) *Multiple imputations for non-response in surveys*. New York: John Wiley and Sons.

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.

Thomas, N., and Gan, N. (1997). Generating multiple imputations for matrix sampling data analyzed with item response models. *Journal of Educational and Behavioral Statistics, 22*, 425-446.

Warm, T. A. (1985). Weighted maximum likelihood estimation of ability in item response theory with tests of finite length. *Technical Report CGI-TR-85-08.* Oklahoma City, OK: U.S. Cost Guard Institute.

Wu, M. L., and Adams, R. J., (2002, April). Plausible Values: Why they are important. Paper presented at the International Objective Measurement workshop, New Orleans, LA.

Wu, M. L., Adams, R. J., and Wilson, M. R. (1997). ConQuest: Multi-aspect test software [Computer program]. Camberwell: Australian Council for Educational Research.