ELSEVIER

Studies in Educational Evaluation 31 (2005) 114-128

Studies in Educational Evaluation

www.elsevier.com/stueduc

THE ROLE OF PLAUSIBLE VALUES IN LARGE-SCALE SURVEYS

Margaret Wu

University of Melbourne, Australia

Abstract

In large-scale assessment programs such as NAEP, TIMSS and PISA, students' achievement data sets provided for secondary analysts contain so-called *plausible values*. Plausible values are multiple imputations of the unobservable latent achievement for each student. In this article it has been shown how plausible values are used to: (1) address concerns with bias in the estimation of certain population parameters when point estimates of latent achievement are used to estimate those population parameters; (2) allow secondary data analysts to employ *standard* techniques and tools (e.g., SPSS, SAS procedures) to analyse achievement data that contains substantial measurement error components; and (3) facilitate the computation of standard errors of estimates when the sample design is complex. The advantages of plausible values have been illustrated by comparing the use of maximum likelihood estimates and plausible values (PV) for estimating a range of population statistics.

Introduction

In large-scale assessments where the purpose is to monitor population standards or progress, the statistics of interest are typically population means, standard deviations, percentages in levels, percentile points, and standard errors associated with these statistics. In this article, these statistics are referred to as population statistics in contrast to statistics relating to individual students, such as individual ability estimates.

0191-491X/04/$ - see front matter
doi: 10.1016/j.stueduc.2005.05.005

Put aside, for a moment, the measurement of student achievement, and suppose that there is an interest in knowing the percentage of people who are over 60 living in a community. If a random sample of people is selected from the community, the percentage, p, of people over 60 in the sample will be an unbiased estimate of the percentage of over 60s in the population. The standard error of the estimate can be computed using the formula $\sqrt{\frac{pq}{n}}$, which is referred to as sampling error.

Now, suppose there is an interest in knowing what percentage of people in this community cannot speak English. While it is relatively easy to determine a person's age, it is not straightforward to determine if each person can speak English, since people's English proficiency is on a continuum: some can speak a few words; some can speak a few sentences; some can speak well enough to be understood, but with many mistakes. So, unlike the measure of people's age, a clear definition of "the ability to speak English" will need to be made. Based on this definition, some assessment will have to be carried out to determine if a person "can speak English". For practical reasons, the assessment can only sample a small part of the English language so as not to place too much burden on each person's time. Consequently, the result of the assessment will contain some uncertainty. This uncertainty is referred to as measurement error. The percentage of people who cannot speak English in the community can be estimated by the percentage, p, of a random sample from the community who failed the assessment. The standard error of this estimate, however, is expected to be larger than $\sqrt{\frac{pq}{n}}$, since there is uncertainty associated with the measurement of each person, in addition to sampling error.

The above example suggests that, if the measurement at the individual level contains error, then this error should be taken into account in the computation of population statistics and their standard errors.

One way to express the degree of uncertainty of measurement at the individual level is to provide several scores for each individual to reflect the magnitude of the error of the individual's estimate. If measurement error is small, then multiple scores for an individual will be close together. If measurement error is large, then multiple scores for an individual will be far apart. These multiple scores for an individual, sometimes known as multiple imputations, are *plausible values.*

The theory and use of plausible values were first developed for the analyses of 1983-84 US National Assessment of Educational Progress (NAEP) data, by Mislevy, Sheehan, Beaton and Johnson (see Mislevy, 1991; Mislevy, Beaton, Kaplan, & Sheehan, 1992; Beaton & Gonzalez, 1995) based on Rubin's (1987) work on multiple imputations. Plausible values were used in all subsequent NAEP surveys, and in surveys such as the Third International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA).

A Formal Definition of Plausible Values

One way to describe plausible values is to say that plausible values represent the range of abilities that a student might reasonably have, given the student's item responses.

There are some differences between plausible values and the θ (student ability parameter) as defined in the usual 1, 2 or 3- parameter logistic item response models. Instead of directly estimating a student's θ, a probability distribution for a student's θ is estimated. That is, instead of obtaining a point estimate for θ, a range of possible values for a student's θ, with an associated probability for each of these values, is estimated. Plausible values are random draws from this (estimated) distribution for a student's θ. This distribution is referred to as the posterior distribution for a student.

Mathematically, the process of drawing plausible values can be described as follows: Given an item response pattern $\mathbf{x}$, and ability θ, let $f(\mathbf{x}|\theta)$ be the item response probability, ($f(\mathbf{x}|\theta)$ could be the 1, 2 or 3-PL model, for example). Further, assume that θ comes from a normal distribution $g(\theta) \sim N(\mu, \sigma^2)$. The function $f(\mathbf{x}|\theta)$ is referred to as the item response model, and $g(\theta)$ the population model.

It can be shown that, the posterior distribution, $h(\theta|\mathbf{x})$, is given by

$$h(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)g(\theta)}{\int f(\mathbf{x}|\theta)g(\theta)d\theta} \qquad \text{(Equation 1)}$$

That is, if a student's item response pattern is $\mathbf{x}$, then the student's posterior (θ) distribution is given by $h(\theta|\mathbf{x})$. Plausible values for a student with item response pattern $\mathbf{x}$ are random draws from the probability distribution with density $h(\theta|\mathbf{x})$. Therefore, plausible values provide not only information about a student's *ability estimate*, but also the uncertainty associated with this estimate.

If many plausible values are drawn from a student's posterior distribution $h(\theta|\mathbf{x})$, these plausible values will form an empirical distribution for $h(\theta|\mathbf{x})$ (as plausible values are observations drawn at random from $h(\theta|\mathbf{x})$). So if a data analyst is given a number of plausible values for each student, the analyst can build an empirical distribution of $h(\theta|\mathbf{x})$ for that student. This approach is adopted because there is no closed form for $h(\theta|\mathbf{x})$ to give to data analysts. Typically, five plausible values are generated for each student, although there does not seem to be strong support in the literature for five. From the simulation results shown in the following sections, it can be seen that, very often, even one plausible value per student is sufficient to adequately recover population parameters.

Various procedures for estimating the parameters of the posterior distribution and for drawing plausible values from the posteriors are given in Beaton (1987), Thomas (1993) and Adams and Wu (2002).

As plausible values are random draws from a student's posterior distribution, plausible values are not appropriate for use as individual student scores for reporting to the students. To illustrate this point, consider the following. Suppose two students have the same raw score on a test, their plausible values are likely to be different as these are random draws from the posterior distribution. Imagine the outcry if two students are given different ability estimates when they have the same response patterns. However, plausible values can be used to estimate population characteristics, and they do a better job than point estimates of abilities.

Point Estimates of Ability Versus Plausible Values

Under the simple Rasch model (Rasch, 1960 [1980]; Wright & Stone, 1979), given a set of item responses from a student, the joint maximum likelihood estimation (JML) (Wright & Stone, 1979) method will produce an ability estimate (MLE) that maximises the likelihood of the observed item responses. One property of the Rasch model is that a student's total score on a test is a sufficient statistic for the ability estimate (Andersen, 1970). That is, only the total score on the test is required, and not the actual response pattern, to compute an ability estimate, provided that all students were administered the same items in the test. Consequently, under JML for the Rasch model, the following observations can be made:

- Every student with the same total score will have the same MLE ability estimate;
- If the maximum score on a test is S, there can only be a maximum of $S+1$ distinct MLE ability estimates for the group of students who took the test;[1]
- If the population distribution is constructed using MLE estimates for individual students, the distribution is a discrete distribution with $S+1$ steps approximating the true population distribution, which is continuous.

Similarly, other point estimates for ability suffer from the same problems as MLE. For example, the procedure using Weighted Maximum Likelihood Estimates (WLE) (Warm, 1985, 1989), while correcting some bias in the JML procedure, still provides one ability estimate for each total score on the test. The Expected A-Posteriori estimate (EAP) (Bock & Aitken, 1981) is the mean of the posterior distribution for each student. Sometimes, EAP can be used as a point estimate for student ability. EAP will also provide the same estimate for students with the same total score, since the posterior distribution is the same for all students with the same total score.

In contrast, when plausible values are used to construct the population distribution, a smoother distribution will be constructed, as students with the same total score, and therefore the same posterior distribution, will likely have different plausible values (random draws from the posterior distribution). The resulting distribution will be a better representation of the underlying continuous population distribution. It should be noted that the distribution built from plausible values from all students is a sample distribution of the population distribution,[2] $g(\theta)$. It follows, therefore, that the sample mean and sample variance of the distribution of plausible values built from all students are unbiased estimates of the population mean and variance of $g(\theta)$ respectively.

In the following section the results have been presented of some comparisons between using point estimates and using plausible values to estimate population mean and variance.

Mean and Variance

It can be shown that if $\hat{\theta}_n$ s are MLEs, the mean of $\hat{\theta}_n$ is an unbiased estimate of μ, the population mean of Θ. But the variance of $\hat{\theta}_n$ is an *over-estimate* of σ^2, the population variance (Mislevy et al., 1992). It can be shown that the mean of EAPs is an unbiased

estimate of the population mean, μ, but the variance of the EAPs is an *under-estimate* of σ^2 (Mislevy et al., 1992). In both (the MLE and EAP) cases, the bias in the variance estimate does not diminish when the sample size increases. But the bias is reduced as the number of items increases.

Simulation Results

Three sets of simulation results are presented below. In the first of them, a data file containing student responses was generated for a 20-item test with 2000 students whose abilities were sampled from $N(0,1)$.[3] The item difficulties were generated from a uniform (-2:2) distribution. WLE, MLE, EAP and five plausible values (PVs) were computed for each student, and the sample mean and variance (across students) were computed for each of these estimates. This process was repeated 100 times (100 replications). The following table gives parameter estimates averaged over the 100 replications. The values in parentheses are the standard deviations of the estimates across the replications—that is the standard errors of the parameter estimates.

Table 1: Mean and Variance Estimates for 20-Item Tests

Estimates averaged over 100 replications	WLE	MLE	EAP	PV1	PV2	PV3	PV4	PV5	Gene-rating value
Estimated population mean of ability with standard error	0.003 (.025)	0.004 (.026)	0.003 (.025)	0.004 (.028)	0.004 (.026)	0.003 (.027)	0.004 (.027)	0.003 (.028)	0
Estimated population variance of ability with standard error	1.300 (.049)	1.456 (.055)	0.778 (.042)	1.002 (.052)	1.006 (.051)	1.000 (.052)	1.004 (.049)	1.005 (.047)	1
	0.98*	1.13*	1.01*						

*Corrected

Table 1 shows that the population mean is estimated well whether WLE, MLE, EAP or PVs are used. None of the estimates is significantly different from the generating value. For the population variance, PVs give estimates closest to the generating value, while WLE, MLE and EAP are all biased. However, WLE-, MLE- and EAP-based estimates recover the generating value reasonably well after a correction factor is applied. The nature of these corrections is discussed in "Reliability as a Measurement Design Effect" (Adams) in this issue of *Studies in Educational Evaluation.*

In the second set of simulations, all parameters are the same as for the first set, except that the population distribution is from $N(2,1)$. That is, the test is *off-target,* as the

average ability of the population is much higher than the average item difficulty. The results are summarised in Table 2.

Table 2: Mean and Variance Estimates for Off-Target 20-Item Tests

Estimates averaged over 100 replications	WLE	MLE	EAP	PV1	PV2	PV3	PV4	PV5	Generating value
Estimated population mean of ability with standard error	1.966 (.031)	2.117 (.031)	2.002 (.032)	2.002 (.035)	2.002 (.033)	2.000 (.033)	2.003 (.032)	2.003 (.035)	2
Estimated population variance of ability with standard error	1.332 (.051)	1.657 (.056)	0.683 (.047)	1.003 (.059)	1.005 (.061)	1.007 (.061)	1.003 (.063)	1.003 (.059)	1
	0.72*	0.89*	1.01*						

* Corrected

Table 2 shows that PVs again recover the generating parameters well. EAP recovers the variance parameter after correction, while WLE and MLE do not recover the generating variance parameter even after correction. These results are not unexpected, when the empirical distributions of the WLE and PV estimates are plotted. The following figure shows the empirical distribution of WLE for one replication:

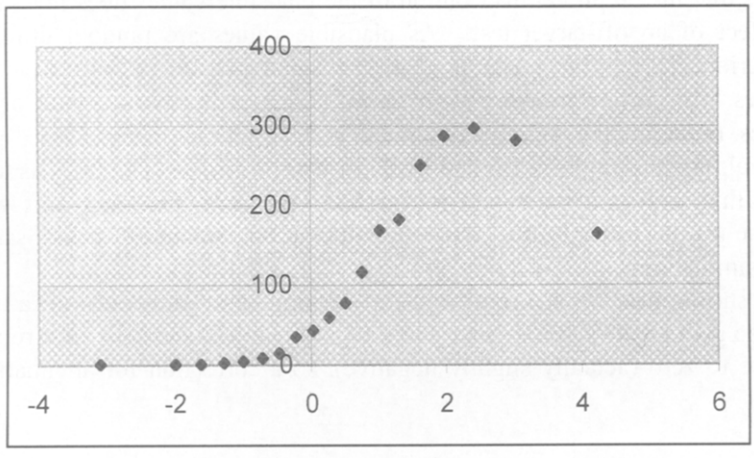

Figure 1: Empirical Distribution of WLE for One Replication

Figure 1 shows a ceiling effect when WLE are computed as ability estimates for an easy test. As raw scores are sufficient statistics for WLE estimates, all students obtaining a perfect score will have the same WLE estimate. In contrast, Figure 2 shows a histogram of plausible values for one replication.

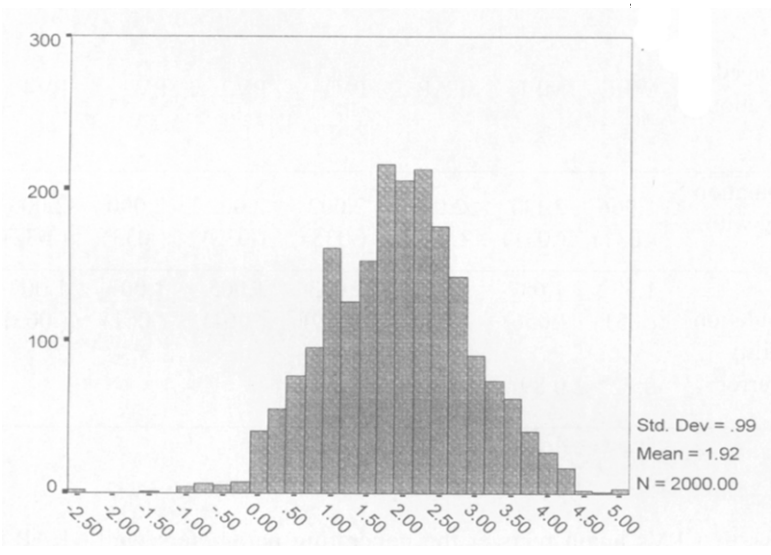

Figure 2: Empirical Distribution of PV for One Replication

Figure 2 shows that the empirical distribution of the plausible values does not suffer from the ceiling effect of an off-target test. As plausible values are random draws from a student's posterior distribution, students with the same raw score will have different plausible values. In particular, those with perfect score will have a range of different plausible values, reflecting the underlying population distribution.

In general, the bias in the WLE and MLE variance estimates increases as test length decreases. A third set of simulations was carried out for a three-item test where the reliability of the test is close to zero. The other simulation parameters are the same as for the first set of simulations.

Table 3 shows that PVs can recover the generating parameters even for a three-item test! EAPs also give good estimates after correction. For WLE and MLE, the reliability of the test is close to zero (actually slightly negative), so a correction for attenuation is not possible.

Table 3: Mean and Variance Estimates for 3-Item Tests

Estimates averaged over 100 replications	WLE	MLE	EAP	PV1	PV2	PV3	PV4	PV5	Generating value
Estimated population mean of ability with standard error	-0.002 (.030)	0.002 (.039)	-0.002 (.036)	0.000 (.041)	-0.004 (.042)	-0.003 (.042)	-0.002 (.041)	-0.003 (.041)	0
Estimated population variance of ability with standard error	1.950 (.263)	2.350 (.178)	0.359 (.061) 0.99*	0.995 (.113)	1.004 (.108)	1.002 (.112)	1.004 (.113)	1.001 (.109)	1

* Corrected

Percent Below Cutpoint and Percentiles

In many system-wide surveys of student achievement, percentages of students achieving a particular level are of interest. In estimating percent in bands and percentiles of population distribution, it can be shown again that plausible values perform better than WLE, MLE and EAP estimates. The following is a simple example showing why point estimates of abilities are not the best for estimating percent in bands or percentiles.

Consider a six-item test, where students' test scores range from 0 to 6. Figure 3 shows the seven (weighted) posterior distributions,[4] corresponding to the seven possible scores, and the corresponding EAP estimates (shown by the black vertical lines).

Suppose the population parameter of interest is the proportion of students below a cut-point, say, −1.0. If EAP is used as the ability estimate, then the proportion of people below −1.0 is the proportion of people obtaining a score of 0. In fact, for any cut-point between EAP_0 and EAP_1 (locations shown by the two left-most black vertical lines), the same proportion is obtained, because the (EAP) ability estimates are discrete, not continuous. The same problem also arises for WLE and MLE estimates. In contrast, the area of the curves of the posterior distributions below −1.0 is a continuous function, and this area contains contributions from all posterior distributions (corresponding to all scores).

As plausible values are random draws from the posterior distributions, the proportion of plausible values below a cut-point provides an estimate of the area of the posterior distributions below that cut-point. The problem associated with the discrete nature of point estimates is overcome by the use of plausible values. Similarly, for percentiles, using plausible values will overcome the problem of having to interpolate between discrete ability estimates.

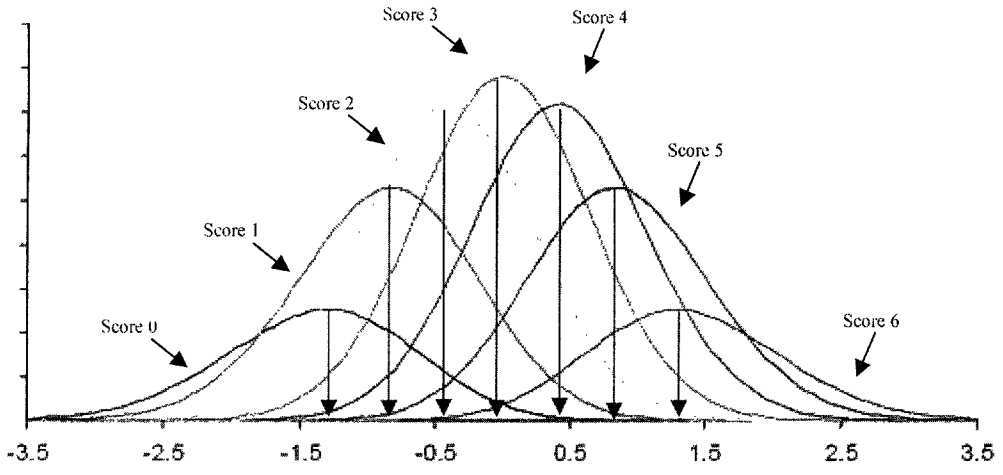

Figure 3: Proficiency on Logit Scale

Simulation Results

In the first set of simulations described above, the 10^{th}, 20^{th}, 30^{th}, 40^{th} and 50^{th} percentile points for the population were estimated. Table 4 shows the results.

The results for PV2, PV3, PV4, PV5 are identical to the results for PV1 (up to the second decimal place). To get a sense of the magnitude of the error in using MLE, the 10^{th} percentile estimate from MLE is in fact around the 7.5^{th} percentile point. While it is expected that there will be a bias for WLE, MLE and EAP, as shown in Table 4, it is more difficult to find the appropriate correction factor for each percentile, as the correction factor is dependent on the percentile. One way is to use $\sqrt{reliability}$ as the correction factor, as shown by the values with asterisks in Table 4. One justification for using this correction factor is as follows. Since the variance of WLE or MLE over-estimates the population variance by a factor equal to the inverse of the reliability, the population standard deviation is over-estimated by $1/\sqrt{reliability}$. That is, the scale built with WLE or MLE is inflated by $1/\sqrt{reliability}$. So it may be reasonable to assume that percentile points are also inflated by $1/\sqrt{reliability}$. Note that in Table 4 the magnitude of the bias in the percentiles fluctuates due to the discrete nature of the ability estimates. Alternatively, if one assumes that the underlying distribution is normal, it will be better to estimate percentiles from WLE, MLE and EAP estimates by first correcting for the variance and then computing percentile estimates from the theorectical normal distribution with the estimated mean and variance. The problem with discrete ability estimates can be overcome in this way.

The simulation results show that, to estimate population statistics, the aggregation of point estimate will often produce biased results. The use of plausible values, on the other hand, produces unbiased results, since the collection of plausible values across students is a sample distribution of the population.

Table 4: Percentile Estimates for 20-Item Tests

Estimates averaged over 100 replications, with standard errors in parentheses	WLE	MLE	EAP	PV1	Generating value
10^{th} percentile	-1.38 (.098) -1.20*	-1.44 (.105) -1.26*	-1.11 (.080) -1.27**	-1.28 (.053)	-1.28
20^{th} percentile	-0.94 (.107) -0.82*	-0.97 (.114) -0.86*	-0.77 (.091) -0.88**	-0.84 (.043)	-0.84
30^{th} percentile	-0.53 (.069) -0.46*	-0.54 (.070) -0.48*	-0.44 (.058) -0.50**	-0.52 (.040)	-0.52
40^{th} percentile	-0.25 (.036) -0.22*	-0.26 (.034) -0.23*	-0.21 (.028) -0.24**	-0.25 (.034)	-0.25
50^{th} percentile	0.00 (.031)	0.00 (.028)	0.00 (.022)	0.00 (.035)	0

* Corrected for attenuation by multiplying the percentile by $\sqrt{reliability}$

** Corrected for attenuation by dividing the percentile by $\sqrt{reliability}$

Conditioning Variables

If there is an interest in estimating statistics for sub-groups of students, such as gender, geographical regions, or language background groups, the generation of plausible values will need to take the group structures into account. Consider an assessment administered to students in two school grades (e.g., TIMSS, Grades 5 and 8). It may be conjectured that, in this case, the underlying population distribution is a mixture of two normal distributions, with different means (μ_1 and μ_2).

In this case, the population distribution is no longer a normal distribution, but a mixture of two normal distributions, which can be specified as

$$g(\theta) \sim N(\mu + \alpha x, \sigma^2) \qquad \text{(Equation 2)}$$

where x = 0 if a student is in group 1 (e.g., Grade 5), and x = 1 if a student is in group 2 (e.g., Grade 8). Note that an assumption is made that the distributions for the two groups have the same variance. In this case, μ, σ^2 and α are estimated. Note also that α is the

difference between the means of the two distributions. That is, group 1 has mean μ (as x = 0), and group 2 has mean $\mu + \alpha$ (as x = 1). σ^2 is the variance of each group (called conditional variance), and not the variance of the whole population. The variable "x" is called a "regressor", or a "conditioning variable", or a "background variable".

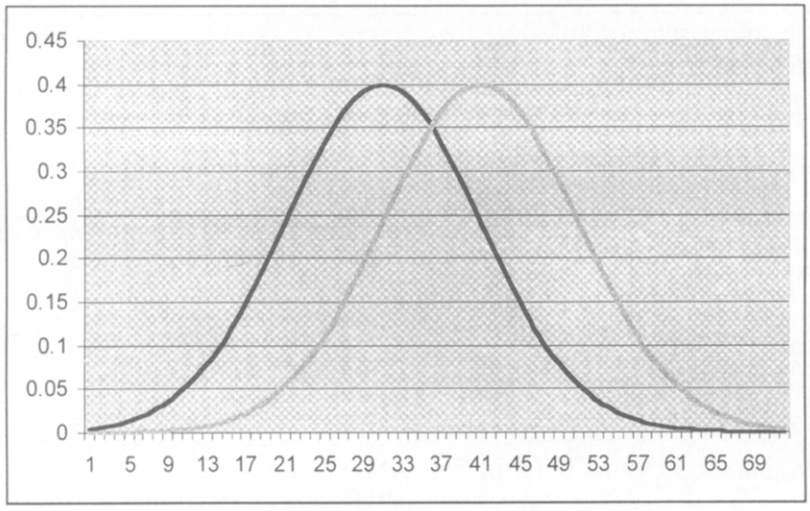

Figure 4: Two Normal Distributions Representing Two Sub-Populations

Latent Regression

More generally, the population model can include many conditioning variables. Equation (2) can be generalised to

$$g(\theta) \sim N(\mu + \alpha x + \beta y + \gamma z + \ldots, \sigma^2) \qquad \text{(Equation 3)}$$

The conditioning variables do not have to be categorical. In fact, most of the conditioning variables are continuous, such as SES. When a conditioning variable is categorical and has more than two categories, dummy variables will need to be constructed as regressors. For example, if there are three school grades (e.g., Grades 5, 8 and 11), then two dummy variables (say, x and y) will need to be constructed. One way to construct the dummy variables is to let x = 0, y = 0 for Grade 3 students; let x = 1, y = 0 for Grade 5 students; and let x = 0, y = 1 for Grade 11 students. In this case, the population distribution can be written as $g(\theta) \sim N(\mu + \alpha x + \beta y, \sigma^2)$. The estimate of μ is the estimated mean for Grade 5; the estimate of $\mu + \alpha$ is the estimated mean for Grade 8; and the estimate of $\mu + \beta$ is the estimated mean for Grade 11.

Why Conditioning Variables Are Needed

To Help Secondary Data Analysts Recover the "Correct" Regression Coefficients

Data analysts are often interested in the relationship between students' background variables and their achievement levels. If regression analyses are carried out with plausible values as dependent variables, and background variables as independent variables, then the "correct" regression coefficients will be "recovered", *if the model that produced the plausible values included the background variables.* If the model that produced the plausible values did not include the background variables as regressors, then the regression coefficients produced with the plausible values will be an under-estimate of the true regression coefficients. The degree of the bias of the regression coefficients will depend on test length and the partial correlation between the variable of interest and the latent variable, after controlling for any conditioning variables that were used. When a regression analysis is run using plausible values generated with a model that did not include the regressors, it is said that a *model mis-specification* has occurred.

Some Commonly Asked Questions About Conditioning

By adding regressors in the population model (e.g., Equation (2)), are spurious differences between the groups actually introduced?

The answer is "no". If there are no differences between the groups, then the estimates of the coefficients $\alpha, \beta, \gamma, \cdots$, will not be significantly different from 0. So by putting a regressor term in the model, it does not necessarily mean that the term has a non-zero effect.

A variant of the above question is:

If plausible values are generated using a model with regressors, isn't it expected that differences between the groups will be found, when secondary analyses use the plausible values? (Argument of circularity)

The answer is "no" again, following the same argument as above. That is, if there is no group difference, then the plausible values generated will be from one population.

One important observation about the marginal maximum likelihood model with latent regression is that two students with the same raw score on a test will get different plausible values (on average) if they are from two different populations. In particular, a student from a population with a lower mean will get lower plausible values (on average) than a student from a population with a higher mean. Consider the following example. Suppose Student A from Grade 5 and Student B from Grade 8 both obtained a raw score of 30 out of 40 on the same test. Since, on average, Grade 5 population mean is lower than Grade 8 population mean, plausible values for Student A will be adjusted "downwards", while plausible values for Student B will be adjusted "upwards".

Using Plausible Values in the Computation of Standard Errors

In large-scale assessments the sampling procedures are often complex. For example, there may be a two-stage sampling process, where schools are first sampled, and then classes or students are sampled within schools. Such sampling procedures call for replication methods such as jackknife or balanced repeated replication for the computation of sampling variance (Wolter, 1985). Given that a collection of plausible values across students forms a sample distribution of the population distribution, replication methods can be applied to each set of plausible values for the computation of sampling variance. However, the size of the measurement error should also be reflected in the standard error for an estimate. As a number of sets of plausible values are typically provided for the data analysts, the variation of the magnitude of the estimate computed using different sets of plausible values provides a variance component for measurement error. In short, the standard errors associated with an estimate will need to be computed from two sources: the variance of plausible values across students will provide the sampling error component, and the variance across sets of plausible values will provide the measurement error component. See Beaton and Gonzalez (1995) for procedures for combining these two sources of error.

Conclusions

In this article an explanation has been provided of what plausible values are, and the rationale for the use of plausible values in surveys where population estimates are the main focus of interest. From simulation results it was seen that plausible values perform well in recovering population parameters such as the mean, variance and percentiles, even when very short tests are administered. Plausible values provide a general methodology that can be used in a systematic way for most population statistics of interest. In contrast, the use of point estimates to construct profiles of population distribution suffers from many problems. While the population mean is unbiased when estimated by the aggregation of point estimates of student abilities, the variance and percentiles are all biased. In a small number of cases, a disattenuation adjustment can be made, but there is no systematic method for adjusting the bias.

The provision of plausible values therefore allows the data analysts to use standard statistical tools to estimate population characteristics. Plausible values are also useful for the computation of standard errors of estimates, particularly when the magnitude of measurement error is large, as is typically the case in large-scale surveys where the focus of interest is population parameters and not individual students.

In closing, it should be pointed out that, while plausible values provide a means for reconstructing the population distribution of interest, the parameters in the model (μ, σ^2, α, β, ...) can be estimated directly. For example, the software ConQuest (Wu, Adams, & Wilson, 1997) provides direct estimation of these parameters. The advantage of this approach is that users will not need to have knowledge of using plausible values, since the software takes care of the estimation procedures. However, the standard errors produced by ConQuest are for simple random samples. If replication methods are required for the computation of sampling variance for complex sampling design, then the use of plausible values is still required.

Notes

1. Assuming, for the moment, there is no difficulty in obtaining an ability estimate for a zero score or a perfect score.
2. Assuming, of course, that the model specification is correct.
3. That is, a normal distribution with a mean of zero and standard deviation of one.
4. The curves in Figure 3 are $h(\theta|x)f(x)=f(x|\theta)g(\theta)$, so they are weighted posteriors. The weighting maintains the shape of each of the posteriors yet shows the relative probability of the different possible scores. Because a Rasch model is used here as the item response model, there is one posterior distribution for each possible score.

References

Adams (this issue of *Studies in Educational Evaluation*) (2005). *Reliability as as measurement design effect.*

Adams, R.J., & Wu, M.L. (Eds.) (2002) *PISA 2000 technical report.* Paris: OECD Publications.

Andersen, E.B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society, 32* (Series B), 283-301.

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

Beaton, A.E., & Gonzalez, E. (1995). *NAEP primer.* Chestnut Hill, MA: Boston College: Boston.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika, 46*, 443-459.

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., & Sheehan, K.M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*, 133–161.

Rasch, G. (1960, 1980) . *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press.

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

Thomas, N. (1993). Asymptotic corrections for multivariate posterior moments with factored likelihood function. *Journal of Computational and Graphical Statistics, 2*, 309-322.

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: U.S. Cost Guard Institute.

Warm, T.A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427-450.

Wolter, K.M. (1985). *Introduction to variance estimation*. New York: Springer-Verlag.

Wright, B.D., & Stone, M.H. (1979). *Best test design*. Chicago: MESA Press.

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

The Author

MARGARET WU has worked as a psychometrician at the Australian Council for Educational Research (ACER) and the University of Melbourne, and also as a consultant to educational institutions within Australia and abroad. She developed a software program ConQuest for the analysis of item response data. This program has the flexibility of fitting a large number of different item response models, and is now widely used both within Australia and internationally. Margaret had a close involvement in large-scale international studies including the Third International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA). She also made substantial contributions to mathematics and problem solving item writing.

Correspondence: <m.wu@unimelb.edu.au>