Institute of Education Sciences
National Center for Education Statistics

NATIONAL INSTITUTE OF STATISTICAL SCIENCES
WORKSHOP REPORT

ACCOUNTING FOR MISSING DATA IN EDUCATION SURVEYS

TABLE OF CONTENTS

NATIONAL INSTITUTE OF STATISTICAL SCIENCES

ACCOUNTING FOR MISSING DATA IN EDUCATION SURVEYS

EXECUTIVE SUMMARY

The National Center for Education Statistics (NCES) charged the National Institute of Statistical Sciences (NISS) with convening a panel of technical experts to consider the issues of accounting for missing data in education surveys. In particular, the panel was asked to address the following questions:

1. Should we analyze and report on datasets for which we have low response rates? What steps should be taken when response rate goals are not met? How should a nonresponse bias study be conducted?
2. How should nonresponse be measured? Should weighting be used in computing response rates? Can the measurement process be made comparable across all surveys? How do we report response rates for surveys involving screening or several rounds of followup? Should we compound conditional response rates? How do we define a complete case? How do we report response rates when nonrespondents are replaced by substitutes?
3. Should NCES generally adopt imputation methods in addition to adjusting for unit nonresponse? Should multiple imputation methods be utilized? What are the cost and practical limitations?
4. Should NCES set minimum response rate standards? If so, should they be the same for future surveys in the planning and design stage? What should they be when addressing public release of an existing data set? Should they be the same in both cases?

Summary and Recommendations

1. Evaluating Nonresponse Bias

Nonresponse bias evaluation should be an integral part of the quality evaluation for all NCES surveys. The extent of the evaluation should be scaled to the seriousness of the nonresponse level based on initial evaluations. Several methods of evaluating nonresponse bias may be employed, ranging from a simple comparison of known characteristics for respondents and nonrespondents to conducting a sample-based followup of nonrespondents on key items. The more intensive methods (followup of nonrespondents) should be implemented when the potential or projected bias is large.

Continue to apply nonresponse adjustment factors at the unit level based on weighting classes, poststratification to known totals, response propensity modeling, or a combination of such techniques, as these are generally effective for reducing nonresponse bias when applied judiciously. For the key items at a minimum, adopt item imputation strategies based on relationships of missing survey characteristics to reported characteristics. Many methods are available for item imputation including matched donor methods (e.g., hot deck) and model-based methods which utilize reported data to predict missing data. Properly conducted, item imputation should also be effective in reducing nonresponse bias. Consider multiple imputation methods to better assess the total error of estimates based on partially imputed data.

2. Measuring Nonresponse

Recognize that the response rate is itself a survey estimate based on the particular sample and the base weights applied to that sample.

Continue to use response rates which incorporate the basic weights at the level of the unit of analysis. Apply base weights at the screening unit level for the screening rate component and base weights at the analysis unit level for the conditional response rate. Express the overall response rate as a product of rates. Technical documentation should include not only the overall response rates, but all unweighted and weighted counts that entered into the computation of each unconditional or conditional response rates.

For the rare cases when matching rather than probability selection approaches are used to substitute for nonrespondents, base the reported response rate on the initial sample only. The response rate for the substitutions should be reported separately to give an indication of the amount of substitution that was used.

If reasonable models for improved imputation of eligibility can be developed, use them to allocate unknown cases to eligible and ineligible categories (an elaboration of Standard 2 of NCES Standard III-02-92).

3. Imputation and Multiple Imputation

Item imputation methods are widely used in government surveys, including NCES surveys.

Continue to use item imputation methods because they can be made effective in reducing nonresponse bias.

In the past, lacking a better alternative, analysts have often treated the imputed values as reported values; however this leads to substantial underestimation of standard errors computed from the data if the amount of missing data is sizeable. Several approaches have been developed and more are being developed to properly estimate the standard errors when data are partially imputed.

The panel is not prepared to recommend a single methodology for NCES to apply routinely, but nonetheless does recommend using a standard error estimation approach which recognizes that data have been imputed.

4. Setting Standards

NCES has taken an important step in developing a Statistical Standards document to guide its statistical activities. These standards should support a process for improving response rates and for improving analytic methods used to deal with nonresponse in all NCES surveys.

There is a danger in setting exact levels of response as a standard because there may be a tendency to be complacent when that level is achieved rather than to strive for continuous improvement in response coverage and the consequent reduction in potential nonresponse bias. Any standards set for individual surveys should be high but within reasonable expectations based on actual experience in similar surveys. A single standard for all surveys does not appear feasible.

PREFACE

The National Center for Education Statistics (NCES) charged the National Institute of Statistical Sciences (NISS) with convening a panel of technical experts to consider the issues of accounting for missing data in education surveys. In particular, the panel was asked to address the following questions:

1. Should we analyze and report on datasets for which we have low response rates? What steps should be taken when response rate goals are not met? How should a nonresponse bias study be conducted?
2. How should nonresponse be measured? Should weighting be used in computing response rates? Can the measurement process be made comparable across all surveys? How do we report response rates for surveys involving screening or several rounds of followup? Should we compound conditional response rates? How do we define a complete case? How do we report response rates when nonrespondents are replaced by substitutes?
3. Should NCES generally adopt imputation methods in addition to adjusting for unit nonresponse? Should multiple imputation methods be utilized? What are the cost and practical limitations?
4. Should NCES set minimum response rate standards? If so, should they be the same for future surveys in the planning and design stage? What should they be when addressing public release of an existing data set? Should they be the same in both cases?

The panel met in-person at NCES and held discussions based on background documents provided by NCES including a comprehensive review of published response counts and response rates for over 30 major NCES surveys spanning from 1988 to the present, extracts from the survey documentation reports, the current NCES statistical standards, and an oral briefing by NCES staff.

NATIONAL INSTITUTE OF STATISTICAL SCIENCES
WORKSHOP REPORT

ACCOUNTING FOR MISSING DATA IN EDUCATION SURVEYS: A WORKSHOP REPORT[1]

I. BACKGROUND FOR PROBLEM DEFINITION

The principal background document for these discussions was a comprehensive review of published response counts and response rates for over 30 major NCES surveys spanning from 1988 to the present[2]. This was supported by extracts from the survey documentation reports. The other major background document was the current NCES Statistical Standards[3]. Further background was provided in an oral briefing by NCES staff, led by Marilyn McMillen.

Initial consideration of three topics identified in the background materials and the briefing were allocated to teams of panelists. Team assignments included both NCES and NISS representatives on each team. Topics for team discussion were defined as follows.

1. *Should we analyze and report on datasets for which we have low response rates? What steps should be taken when response rate goals are not met? How should a nonresponse bias study be conducted?*
2. *How should nonresponse be measured? Should weighting be used in computing response rates? Can the measurement process be made comparable across all surveys? How do we report response rates for surveys involving screening or several rounds of followup? Should we compound conditional response rates? How do we define a complete case? How do we report response rates when nonrespondents are replaced by substitutes?*
3. *Should NCES generally adopt imputation methods in addition to adjusting for unit nonresponse? Should multiple imputation methods be utilized? What are the cost and practical limitations?*

An additional item was not assigned to a team, but arose in the process of discussing response measurement: *Should NCES set minimum response rate standards? If so, should they be the same for future surveys in the planning and design stage? What should they be when addressing public release of an existing data set? Should they be the same in both cases?*

[1] Technical Report Number 90.
[2] Marilyn McMillen memorandum of August 27, 1998.
[3] NCES Statistical Standards (NCES 92-02lr) June 1992.

II. DISCUSSION

The presentation here follows the topical areas identified above and does not necessarily reflect the order of discussion.

2.1 Nonresponse Bias

Nonresponse bias must be viewed in the context of total survey error. The mean squared error measure captures both sampling error and nonsampling errors in surveys. It is expressed as

$$\text{MSE} = \text{Variance} + \text{Bias}^2.$$

Nonresponse bias is only one component of total bias. Bias in general takes on more significance when the sampling error (square root of the variance of the estimate) is small, because then it becomes a proportionately larger component of total error. For major NCES surveys where the sample size is large with corresponding low sampling error, bias becomes an important component of total error. Nonresponse bias can then be a major component of total bias (and of total error) when response rates are low.

Nonresponse bias, B_{nr}, for a population mean is the product of two components: (1) the response rate, r, and (2) the difference, d, between the expected values of estimates for respondents and nonrespondents; i.e.,

$$B_{nr} = (1 - r)\, d.$$

Note that $d = E(e_r) - E(e_{nr})$, or the expected value of the estimate for respondents minus the expected value of the estimate for nonrespondents (assuming nonrespondents would respond). It should also be noted that if the survey involves unequal weighting, the bias expression is actually a function of the weighted response rate with appropriate weights for the entire sample of respondents and nonrespondents. This expression of the bias also assumes that no bias-reducing weight adjustment procedures, such as weighting class adjustments, response propensity models (Rosenbaum & Rubin 1983; Little 1986, 1988), or post-stratification to known totals, have been employed. Expressions of Nonresponse bias for other parameters are different. For example, discarding the incomplete cases yields unbiased estimates for regression parameters if missingness depends on the covariates, although it yields biased estimates of means or proportions (Little 1997).

This formula shows us that the surest way to control nonresponse bias is to maintain a high response rate. When a survey is completed, the response rate is known and the estimate for respondents is known. Because the estimate for nonrespondents is not known, the value of the difference, d, is also not known, but keeping the response rate high ensures that the bias will remain small.

We can also relate bias to mean squared error. The bias due to nonresponse was expressed above as

$$B = (1 - r)\, d$$

and the mean squared error as

$$MSE = B2 + V.$$

The relative bias (relative to total error) can then be expressed as

$$relbias = \frac{B}{\sqrt{B^2+V}} = \frac{1}{\sqrt{1+\frac{V}{B^2}}} = \frac{1}{\sqrt{1+\frac{V}{(1-r)^2\ d^2}}}.$$

From this expression, it can be seen that the relative bias is a monotonic function of

$$\frac{1-r}{\sqrt{V}} \propto (1-r)\ \sqrt{n}$$

because the variance is approximately a constant divided by the sample size, n. The relative impact of bias on total mean squared error increases as the response rate decreases and/or as the sample size increases.

Because we know the response rate and we have an estimate of the population parameter based on respondents, the only term needed to evaluate the bias is a good estimate of the survey characteristics for nonrespondents. Several methods of conducting nonresponse bias studies were discussed:

Method A: Compare respondents and nonrespondents on observed characteristics. This gives clues about whether bias is present, but may be limited by the absences of observed characteristics sufficiently related to the survey outcomes.

Method B: In surveys that involve successive levels of effort to reduce nonresponse, consider plotting the estimates as the level of effort increases. If a trend is established, it may be possible to extrapolate to a measure of total bias.

Method C: Conduct a followup study of nonrespondents on key variables. Use more intensive methods, incentives, etc., along with reduced burden to obtain as a high a response rate as possible for key measures.

Method D: Impute for missing values, based on an estimated predictive distribution given the known variables. Multiple imputation provides estimates of imputation uncertainty.

Method E: Conduct sensitivity analyses to assess the effect of posited differences between survey respondents and nonrespondents on survey estimates. (See, for example, Rubin 1977, Little & Wang 1996.)

The five methods listed are not mutually exclusive. Some methods, particularly Methods B and C could be more powerful when used in tandem, e.g. the data from a nonrespondents survey could be used to adjust the trend line based on late respondents.

What do we do when unit and/or item response rates are very low? Generally, it seems silly to throw away any data if retaining the data advances our knowledge. Imputation methods must assume a model and we know every model is wrong. However, because imputation models only affect the predictions of the missing values, the model misspecification is confined to the impact on those predictions; therefore, the impact of model misspecification is less when the amount of missing data is small than when the amount of missing data is extensive.

The convention for panel research such as the Survey of Income and Program Participation (SIPP) has been to employ wave weights using all unit responses from each wave and special panel weights for longitudinal analysis which define unit response rates in terms of responses to all waves. SIPP now imputes for missing wave data, provided that the preceding and succeeding waves to the missing wave are not missing.

A major problem with any imputation effort is preserving all interrelations among variables. We want to assume that missing variables are highly correlated with nonmissing variables and the probability of being missing does not depend on this correlation.

Weighting for nonresponse controls bias, but may increase variance. Correct application of imputation controls bias and may also decrease variance.

Typically, a flat (uninformative) prior is used in Bayesian approaches in order to limit subjectivity and maintain comparability with standard frequentist answers when data are complete. The "real" prior is, however, the predictive model. The assumption that data are missing at random, given the covariates, is key to many applications, although models that do not assume that data are missing at random have been applied in some areas, such as attrition in longitudinal data (Little 1995).

If NCES wishes to test multiple imputation methods on current datasets, they should select datasets with informative covariates. It should be practicable to apply multiple imputation methods even with a large number of variables with different proportions missing. It would be necessary to develop models to impute the missing variables and to accustom analysts to using the appropriate analytic techniques with imputed data.

Without proper care, inconsistencies can be generated as a result of imputation. Gibbs sampling approaches and proper conditioning can resolve the inconsistency problem, but may require intensive efforts and high costs.

An alternative to imputation that does not involve discarding data is to use a method that accepts data in a non-rectangular form, such as maximum likelihood methods. These methods are becoming increasingly popular for the analysis of repeated measures data, using software such as SAS Proc Mixed (SAS 1992). Analytic techniques for nonignorable nonresponse in repeated measures data are reviewed in Little (1995). Econometric methods have also been applied (Heckman 1976).

The choices facing NCES with existing data sets are: (1) Report analyses based on complete cases, perhaps weighted to make them more representative of the full sample. (2) Impute for missing values. (3) Don't report if the response rate is low. If item response is high, imputation won't hurt; if the item response is low, multiple imputation can help. After imputation, estimates from the complete case analysis can be compared to the imputed case analysis for differences in estimates (bias) and variances of estimates (comparing multiple imputation variance estimates with the variance estimate for complete cases).

Response analysis comparing characteristics of respondents and nonrespondents is relevant to the decision to employ imputation techniques. Covariates address bias. If the distributions of respondents and nonrespondents are similar, imputation will not reduce the bias. Imputation may increase precision if the observed covariates are good predictors of the missing values. Weighting for nonresponse without trimming can increase variance with some reduction in bias, whereas weighting with trimming can control variance while also reducing bias. (See Little et al. 1997.)

Cost may be an issue in applying extensive imputation schemes. NCES should review the experience of other agencies (e.g., NCHS, StatCanada, others) that have implemented imputation and multiple imputation approaches.

It is difficult to anticipate all future analyses in planning effective imputation methods. For existing datasets, this may be less of a problem because key reporting subgroups have already been identified by NCES.

Imputation modeling often uncovers data editing problems.

2.2 Nonresponse Computation

Several issues were addressed in the discussion including (1) use of sampling weights in computing response rates; (2) multiplication of successive response rates in multi-wave surveys or surveys that involve a screening interview; (3) the definition of a respondent; (4) item response rates; (5) documentation of response rate calculation; (6) treatment of substitutes; and (7) treatment of a survey response rate as an estimate.

Standard 1 of NCES standard III-02-92 prescribes the use of base weights in computing response rates. We interpret base weights to be weights that have not been in any way adjusted for nonresponse; these weights are sometimes called the design-based weights. Standard 1 further states that "When the sampling unit is not the unit of analysis, it is appropriate to multiply the sampling weight of the sampling unit by the sampling weight of the unit of analysis." Standard 2 prescribes removal of "weighted out-of-scope units" from the denominator and suggests imputing the number of eligible and ineligible units among the "unable to contact" (unknown eligibility) units. The response rate is then computed as:

$$r = \frac{\textit{weighted number of completed interviews}}{\textit{weighted number of units sampled - weighted number of out - of - scope units}}.$$

This can be compared to the Council of American Survey Research Organizations (CASRO) suggested response rate formula[4] which prescribes a specific way to impute the number of out-of-scope units:

$$\textit{response rate} = \frac{\textit{completed}}{\textit{eligible} + \dfrac{\textit{eligible}}{\textit{eligible} + \textit{ineligible}} * \textit{unknown}}.$$

We recommend the use of weighted response rates at the unit of analysis level as prescribed in the NCES standards. Weighted response rates more realistically portray the potential magnitude of the nonresponse bias, B_{nr}, as illustrated in the discussion of nonresponse bias above. Note that when probability-proportional-to-size sampling is used uniformly throughout the sample design and the size measures are proportional to the number of analysis units per sampling unit, the weighted and unweighted response rates are equivalent. (We temporarily defer comment on the method of imputation for counts of eligible units from among units with unknown eligibility.)

Multiplication of rates becomes an issue when a screening questionnaire is applied at the sampling unit level to determine eligibility or to enumerate all the analysis units associated with a sampling unit. For example, a sample of addresses could be screened for occupied households. At each occupied household, the screening questionnaire might be designed to enumerate all permanent residents, all children of school age, or all owned computers utilized by members of the household. A sample of analysis units could then be selected from the enumerated analysis units at each household. A more general form of the CASRO formula can then be written as

[4] The CASRO formula can be found on the World Wide Web at http://home.clara.net/sisa/resphlp.htm.

$$overall\ response\ rate = \frac{eligible + ineligible}{eligible + ineligible + unknown} * \frac{interviewed}{eligible}$$

$$= screening\ rate\ *\ conditional\ interview\ response\ rate.$$

If there is no more than one analysis unit per sampling unit, this formula reduces (algebraically) to the CASRO formula and, with appropriate imputation of the unknowns, to the NCES formula assuming equal weighting. When this is not the case, the formula provides a basis for computing an overall response rate as the product of an initial-stage response rate and one or more conditional response rates for the subsequent stage or stages.

We recommend applying base weights at the screening unit level for the screening rate component and base weights at the analysis unit level for the conditional response rate, and expressing the overall response rate as a product of rates. We further recommend that the individual rates, and not just their product, be included in the survey documentation.

We now return to the issue of alternative methods of imputing eligibility status for sampling units not fully screened. There may be some cases when, although the eligibility status of a sampling unit is not completely known, other data are available that provide some indication of the probability of the unit actually being eligible. As an example, in telephone surveys, the call history may indicate no contact at all after several calls for some households and the lack of a qualified adult respondent for other households. If reasonable methods can be developed for modeling the probability of eligibility for these two types of cases, that information could be used to improve over the proportional allocation of unknowns based on completed cases implied by the formulas above. (The problem of unknown eligibility is in some ways very much like that of unknown values for nonrespondents. Some of the same approaches might be used to estimate eligibility as are used for estimating data values for nonrespondents. In particular, a supplemental data collection effort aimed at the units of unknown eligibility could be conducted; such data could aid in further modeling.)

Before response rates may be computed, it is necessary to examine and edit each observation and determine whether enough information has been provided to include it in the data file. Any item included in the data file should be treated as a unit-level response. Some cutoff must be established regarding which items must be answered or what proportion of key items must be answered in order to consider an observation sufficiently complete to include in analyses. The choice of cutoff point determines the boundary between unit nonresponse and item nonresponse and should be guided primarily by the planned analyses. Observations just above the cutoff point will suffer from low item response rates and may require heavy imputation for missing values.

Technical reports of NCES surveys have varied in the amount of supporting detail provided for the reported response rates. We recommend complete documentation of all response rate calculations. Documentation should include the formulas used as well as inputs to the formulas. In addition, we recommend provision of both unweighted counts and weighted counts for total sample, sample determined eligible, sample determined ineligible, and sample with eligibility unknown. If screening is employed and the overall response rate is computed as a product of unconditional and conditional response rates, similar unweighted and weighted counts should be provided for the computation of response rate in the product formula. In addition, it is essential that the documentation include detailed definitions of various sample (call) result categories. Ambiguity about what types of sample results a category represents makes it

difficult to interpret reported rates. It would be useful for some steps to be taken to ensure that these definitions are uniform across NCES surveys.

Substitution is used in only limited cases in NCES surveys. More generally, the initial sample is specified at an adequate size to yield the desired number of respondents based on projected screening yields and projected response rates. For some surveys, supplemental probability subsamples may be held in abeyance and released only if projected yields based on partial response data indicate a shortfall in the final sample size. Neither of these two cases requires any special approach to computing response rates; all cases released for data collection may be treated as the base sample with weights defined to reflect that particular sample size and the particular sampling units included in the sample.

In some surveys (e.g., NAEP multi-stage selection of schools), substitution for school refusals based on matched schools is used to preserve the sample within primary sampling units. The matching approach is used to preserve the characteristics of the initial sample and is viewed by some as being akin to imputation for missing values. Several options exist for computing the response rate when substitutes are used: (1) Base the response rate calculation on the initial sample only and totally ignore the substitutes in both the numerator and denominator (current NAEP practice); (2) Add substitutes to the numerator in the response rate; and (3) Add substitutes to both the numerator and the denominator. Method 2 was deemed clearly inappropriate. Method 3 might make sense if a substitute were selected at random within a stratum, but even then applying weights becomes problematic in many cases. For the rare cases when matching rather than probability selection approaches are used to substitute for nonrespondents, we recommend basing the reported response rate on the initial sample only.

Item response rates should be reported as well as unit response rates. There was some discussion of reporting an overall, single response rate based on the ratio of the count of items actually answered to the count of items that should have been answered. Such a single number would allow comparison of the quality of survey item response for the responding units across surveys in conjunction with their overall unit response rates. However, if gate questions are left unanswered, the concept of "items that should have been answered" may be ambiguous or require imputation. Additionally, the combined measure would treat all items equally and might not properly account for critical items. We recommend reporting item response rates individually with a focus on critical items or sets of items. The overall response rate for an item could then be computed as the product of the overall unit response rate times the item response rate.

Treatment of computed survey response rates as estimates was discussed. This approach makes sense for future planning in that response rates will not be replicated exactly even when the same protocol is followed. For the current or completed survey, the ultimate use of the response rate is to estimate the bias or possible limits on the bias under various assumptions about the differences between respondents and nonrespondents, and for this purpose it also seems appropriate to treat the response rate as an estimate. For the purpose of managing contractors and subcontractors, the concept of a confidence interval around the estimated response rate might be useful and appropriate; this would provide greater leeway for response rates in small surveys or in small subpopulations within larger surveys to meet a specified response rate goal.

2.3 Imputation Issues

An overview of multiple imputation methods was presented by Rod Little. Single imputation methods can seriously understate uncertainty, yielding confidence intervals that are too narrow and p-values that are

too small (Rubin & Schenker 1986). Multiple imputation (Rubin 1987, 1996; Rubin & Schenker 1986, Little 1988) addresses this problem by allowing imputation uncertainty to be incorporated into the analysis. Multiple imputation also provides simple estimates of the loss of information (that is, the added variance of estimates) caused by the fact that values are missing. Multiple imputation under alternative models can be used to assess sensitivity to various imputation models, allowing the user to be more confident about the conclusions drawn from the data.

Multiple imputation methods are gaining acceptability and are already being used by NCES in the guise of NAEP plausible values. For recent applications to a large government survey, see Khare et al. (1993), and Ezzati-Rice et al. (1995). Multiple imputation techniques require models for the joint distribution of the survey variables, with appropriate attention to data transformations, etc. Therefore, if the set of survey variables is very large the modeling effort may need to be restricted to a set of key variables of interest. If about five complete data sets are imputed, available software can be used to obtain parameter estimates and variance estimates. An additional simple routine (e.g., a SAS macro) is then required to compute the average parameter estimate, the average within-imputation variance, and the additional variance contribution associated with multiple imputations. Software for performing multiple imputation includes Schafer's algorithms, currently available at his web site but also being incorporated in S-Splus (Schafer 1996), and the Bayesian simulation software, Bayes Using Gibbs' Sampling (BUGS). For longitudinal data the recent software package, SOLAS, creates multiple imputations based on a simple model focused on bias reduction.

Multiple imputation theory is closely related to Bayesian simulation methods such as the Gibbs' sampler (Gelfand & Smith 1990; Tanner 1991). The data can be viewed as consisting of complete observations and observations with missing data, $Y = (Y_{obs}, Y_{mis})$. Multiple imputations are generated from some predictive distribution (e.g., regression, hot deck, etc.). Rubin (1987) distinguishes between improper and proper multiple imputation (see also Fay 1992); for the regression case, proper imputation requires drawing from the regression parameter distribution as well as from the error distribution, and more fully captures the total imputation uncertainty. Suppose M imputations are performed for each missing item using a proper multiple imputation method; i.e., for $j = 1, 2, \ldots, M$ complete data sets are created from the predictive distribution and designated by $Y^{(j)} = (Y_{obs}, Y^{(j)}{}_{mis})$. An estimate, $\hat{\theta}^{(j)}$, of some population parameter, θ, can then be generated from each complete data set. The multiple imputation estimate of θ is the simple average over the multiple imputations:

$$\hat{\theta}_{MI} = \frac{\sum_{j=1}^{M} \hat{\theta}^{(j)}}{M}.$$

Its variance is approximated by

$$Var(\hat{\theta}_{MI}) \approx \frac{1}{M}\sum_{j=1}^{M} Var(\hat{\theta}^{(j)}) + \frac{M+1}{M}\sum_{j=1}^{M} \frac{(\hat{\theta}^{(j)} - \hat{\theta}_{MI})^2}{M-1}.$$

If the predictive model is properly specified, variance estimates are consistent. The first variance component in the above equation is the average within-imputation variance, W; the second component is the between-imputations component, B. An expression for the fraction of missing data is then

$$\gamma = \frac{B}{B + W}.$$

The fraction of missing data depends on the parameters that are estimated and on how much information is used for imputation. With no covariates and simple random sampling, we note that $W \propto 1/n$ where n is the total sample size, and that $W + B \propto 1/m$ where m is the complete sample size, the fraction of missing data reduces to the nonresponse rate

$$\gamma = \frac{n - m}{n}.$$

Multiple imputation is a very general method in that it provides valid inferences for any target parameter (i.e., overall means, subclass means, regression coefficients, correlations, etc.). The method requires considerable resources to generate good predictive distributions of the missing values in large survey settings, but the key idea is that the effort to create the imputations is carried out just once by the data provider, and does not need to be replicated by every user of the data. The creation of a single set of multiple imputations by the data provider promotes consistency of analyses because it avoids anomalies created when different missing-data adjustments are employed by different users of the same database.

Bootstrapping methods can also be used to estimate the added variance from imputation, and adjustment methods have been proposed to assess imputation variance for particular survey estimands. For a discussion of these alternatives, see the papers by Fay (1996) and Rao (1996) that accompany the review of multiple imputation by Rubin (1996), and the accompanying discussions of these three papers.

Graham Kalton pointed out that other approaches are being developed for taking account of imputed data in variance estimation using balanced repeated replications (Shao, Chen, & Yinzhong 1998),jackknife replications (Rao & Shao 1992, Kovar & Chen 1994), bootstrapping (Shao & Sitter 1996), and Taylor's series approaches (Rao 1996, Sarndal 1992, Fay 1996). A limitation of these alternative methods at present is that they are confined to simple estimates like means and totals. Hopefully, ongoing research will address this limitation.

2.4 Response Rate Standards

NCES has established standards for the design of cross-sectional and longitudinal surveys (NCES Standard 1-02-92). Minimum response rates have also been set for some surveys (e.g., 70 percent overall unit response rates for state assessment surveys) as a basis for including the survey in a national report. Generally, it appears that standards for release of data may be lower than those used in planning the survey. The Office of Management and Budget sets fairly rigorous response rate standards for the advance planning of surveys, but has not routinely prohibited the production of reports when the actual response rates do not meet the survey design standards. Often response rates for subpopulations reported separately in analytic reports are much lower than the average response rate for the entire survey.

We generally agreed that if design standards are to be set, they should be high. Specifying low targeted response rates would almost surely lead to response rates no higher than the target, especially in the

competitive bidding environment prevalent for many NCES surveys. The principles of Deming would encourage a process of continuous improvement rather than simple standard setting.

Nonresponse bias is only one component of total survey error. Any standards set should recognize all contributions to error. In particular, tradeoffs between unit and item Nonresponse need to be taken into account. In addition, costs must be balanced against potential gains in survey quality.

We noted that other government statistical agencies vary on their setting of standards. The National Center for Health Statistics and the Energy Information Administration have written standards; the Bureau of the Census does not.

III. CLOSING SESSION COMMENTS (Much of what was said here is a repetition of earlier discussions.)

- Weighting methods are reasonable for handling unit nonresponse where covariate information on the nonrespondents is limited. Imputation is the better approach for item nonresponse where the pattern of nonresponse is complex, and the covariate information available to predict the missing values is extensive. With repeated surveys, weighting is often used in practice, although this method can be highly inefficient, particularly for cases with information recorded both before and after a missing wave. Maximum likelihood methods such as those in SAS Proc Mixed are suggested for repeated measures analyses of key survey outcomes. For public use files, imputation is also recommended, limited to major survey estimands if the set of variables to be analyzed is too extensive to be manageable given available resources. When imputation is carried out, multiple imputation is recommended to allow the impact of imputation uncertainty to be measured and incorporated in final inferences.
- Multiple imputation is an analog to the NAEP procedure for generating plausible values.
- Adoption of multiple imputation strategies by NCES might require modifications to existing contracts.
- NCES should implement routine evaluations of nonresponse.
- NCES should evaluate the experience of Stat Canada, some of the Scandinavian countries, and other U.S. government agencies.
- First efforts should address studies with low response rates.
- NCES should investigate the use of external data for poststratification and raking adjustment purposes.
- Expectations must be realistic. It may be necessary to set different standards for cross- sectional and longitudinal studies. If data from other panels are used for imputation (or weighting adjustment), compounded nonresponse is a less serious problem.
- Moderate nonresponse may be ignorable for change measures and regression parameters even when it is nonignorable for estimating means or proportions.

IV. SUMMARY AND RECOMMENDATIONS:

4.1 Evaluating Nonresponse Bias

A nonresponse bias evaluation should be integral part of the quality evaluation of all NCES surveys. The extent of the evaluation should be scaled to the seriousness of the nonresponse level based on initial evaluations. Several methods of evaluating nonresponse bias may be employed, ranging from a simple comparison of known characteristics for respondents and nonrespondents to conducting a sample-based followup of nonrespondents on key items. The more intensive methods (followup of nonrespondents) should be implemented when the potential or projected bias is large.

Continue to apply nonresponse adjustment factors at the unit level based on weighting classes, poststratification to known totals, response propensity modeling, or a combination of such techniques, as these are generally effective for reducing nonresponse bias when applied judiciously. For at least the key items, adopt item imputation strategies based on relationships of missing survey characteristics to reported characteristics. Many methods are available for item imputation including matched donor methods (e.g., hot deck) and model-based methods which utilize reported data to predict missing data. Properly conducted, item imputation should also be effective in reducing nonresponse bias. Consider multiple imputation methods to better assess the total error of estimates based on partially imputed data.

4.2 Measuring Nonresponse

Recognize that the response rate is itself a survey estimate based on the particular sample and the base weights applied to that sample.

Continue to use response rates which incorporate the basic weights at the level of the unit of analysis. Apply base weights at the screening unit level for the screening rate component and base weights at the analysis unit level for the conditional response rate. Express the overall response rate as a product of rates. Technical documentation should include not only the overall response rates, but all unweighted and weighted counts that entered into the computation of each unconditional or conditional response rates.

For the rare cases when matching rather than probability selection approaches are used to substitute for nonrespondents, base the reported response rate on the initial sample only. The response rate for the substitutions should be reported separately to give an indication of the amount of substitution that was used.

If reasonable models for improved imputation of eligibility can be developed, use them to allocate unknown cases to eligible and ineligible categories (an elaboration of Standard 2 of NCES Standard III-02-92).

4.3 Imputation and Multiple Imputation

Item imputation methods are widely used in government surveys, including NCES surveys. We recommend the continued use of item imputation methods because we agree that they can be made effective in reducing nonresponse bias.

In the past, lacking a better alternative, analysts have often treated the imputed values as reported values; this leads to substantial underestimation of standard errors computed from the data if the amount of

missing data is sizeable. Several approaches have been developed and more are being developed to properly estimate the standard errors when data are partially imputed.

We are not prepared to recommend a single methodology for NCES to apply routinely, but we do recommend using a standard error estimation approach which recognizes that data have been imputed.

4.4 Setting Standards

NCES has taken an important step in developing a Statistical Standards document to guide its statistical activities. These standards should support a process for improving response rates and for improving analytic methods used to deal with nonresponse in all NCES surveys.

There is a danger in setting exact levels of response as a standard because there may be a tendency to be complacent when that level is achieved rather than to strive for continuous improvement in response coverage and the consequent reduction in potential nonresponse bias. Any standards set for individual surveys should be high but within reasonable expectations based on actual experience in similar surveys. A single standard for all surveys does not appear feasible.

APPENDICES

Appendix A: References

Ezzati-Rice, T., Johnson, W., Khare, M., Little, R.J.A., Rubin, D., and Schafer, J.L. (1995). A Simulation Study to Evaluate the Performance of Model-Based Multiple Imputations in NCHS Health Examination Surveys. *Proceedings of the 1995 Annual Research Conference, US. Bureau of the Census,* 257-266.

Fay, R.E. (1992). When are Inferences from Multiple Imputation Valid? *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 227-232.

Fay, R.E. (1996). Alternative Paradigms for the Analysis of Imputed Survey Data. *Journal of the American Statistical Association, 91,* 490-498.

Gelfand, A.E., and Smith, A.F.M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association, 85,* 398-409.

Heckman, J. (1976). The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. *Annals of Economic and Social Measurement, 5,* 475-492.

Khare, M., Little, R.J.A., Rubin, D.B., and Schafer, J.L. (1993). Multiple Imputation of NHANES III. *Proceedings of the Survey Research Methods Section, American Statistical Association.*

Kovar, J.G., and Chen, E.J. (1994). Jackknife Variance Estimation of Imputed Survey Data. *Survey Methodology, 20,* 45-52.

Little, R.J.A. (1986). Survey Nonresponse Adjustments. *International Statistical Review, 54,* 139-157.

Little, R.J.A. (1988). Missing Data Adjustments in Large Surveys. *Journal of Business and Economic Statistics, 6,* 287-301.

Little, R.J.A. (1995). Modeling the Drop-Out Mechanism in Longitudinal Studies. *Journal of the American Statistical Association, 90,* 1112-1121.

Little, R.J.A. (1997). Biostatistical Analysis with Missing Data. In P. Armitage and T. Colton (Eds.), *Encyclopedia of Biostatistics.* London: Wiley.

Little, R.J.A., Lewitzky, S., Heeringa, S., Lepkowski, J., and Kessler, R.C. (1997). An Assessment of Weighting Methodology for the National Comorbidity Study. *American Journal of Epidemiology, 146,* 439-449.

Little, R.J.A., and Wang, Y.-X. (1996). Pattern-Mixture Models for Multivariate Incomplete Data with Covariates. *Biometrics, 52,* 98-111.

Rao, J.N.K. (1996). On Variance Estimation with Imputed Survey Data. *Journal of the American Statistical Association, 91,* 499-506 (with discussion).

Rao, J.N.K., and Shao, J. (1992). Jackknife Variance Estimation with Survey Data Under Hot Deck Imputation. *Biometrika, 79,* 811-822.

Rosenbaum, P.R., and Rubin, D.B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika, 70,* 41-55.

Rubin, D.B. (1977). Formalizing Subjective Notions about the Effect of Nonrespondents in Sample Surveys. *Journal of the American Statistical Association, 72,* 538-543.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley. Rubin, D.B. (1996). Multiple Imputation after 18+ Years. *Journal of the American Statistical Association*, 91, 473-489.

Rubin, D.B., and Schenker, N. (1986). Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse. *Journal of the American Statistical Association, 81,* 366-374.

SAS. (1992). *The Mixed Procedure, in SAS/STAT Software: Changes and Enhancements, Release 6.07* (Technical Report P-229). Cary NC: SAS Institute, Inc.

Sarndal, C.E. (1992). Methods for Estimating the Precision of Survey Estimates when Imputation has been used. *Survey Methodology, 18,* 241-252.

Schafer, J.L. (1996). *Analysis of Incomplete Multivariate Data.* London: Chapman & Hall.

Shao, J., Chen Y., and Yinzhong C. (1998). Balanced Repeated Replication for Stratified Multistage Survey Data under Imputation. *Journal of the American Statistical Association, 93,* 819-831.

Shao, J., and Sitter R.R. (1996). Bootstrap for Imputed Survey Data. *Journal of the American Statistical Association, 91,* 1278-1288.

Tanner, M.A. (1991). *Tools for Statistical Inference: Observed Data and Data Augmentation Methods.* New York: Springer-Verlag.

Appendix B: Expert Workshop Participants

Workshop Participants

Johnny Blair - University of Maryland

James Chromy - Research Triangle Institute

Richard Jaeger - University of North Carolina at Greensboro

Lyle V. Jones - University of North Carolina at Chapel Hill

Graham Kalton - Westat

Roderick Little - University of Michigan

Ingram Olkin - Stanford University

Valerie S. L. Williams - North Carolina Central University

National Center for Education Statistics

Dennis Carroll

Pascal Forgione

Daniel Kasprzyk

Marilyn McMillen

Martin Orland

Gary Phillips

Education Statistics Services Institute

Karol Krotki

Workshop convened by National Institute of Statistical Sciences

Jerome Sacks - National Institute of Statistical Sciences