Institute of Education Sciences
National Center for Education Statistics

NATIONAL INSTITUTE OF STATISTICAL SCIENCES
TECHNICAL EXPERT PANEL REPORT

STUDY DESIGN DECISIONS FOR POSTSECONDARY SAMPLE SURVEYS

TABLE OF CONTENTS

NATIONAL INSTITUTE OF STATISTICAL SCIENCES

TECHNICAL EXPERT PANEL REPORT ON STUDY DESIGN DECISIONS FOR NATIONAL POSTSECONDARY SAMPLE SURVEYS

EXECUTIVE SUMMARY

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 for design of the National Postsecondary Student Aid Study. In particular, the panel was asked to address the sample design strategies currently in use for NPSAS, strategies including responsive designs employed by other recent NCES surveys, and applicable innovations, methodological advances and also alternative design approaches from other large-scale federal or governmental agency surveys and assessments. Broad topics to be considered included: i) Sample Design, Sampling Error and Design Effects, ii) Measurement Error, and iii) Imputation and Administrative Data.

The panel met first via teleconference to discuss materials prepared by NCES staff and then met in person with presentations and discussions with NCES staff. The panel held further closed teleconferences to prepare this report.

Study Design Decisions for National Postsecondary Sample Surveys

BACKGROUND

As the primary federal entity for collecting and analyzing data related to education in the United States, the National Center for Education Statistics (NCES) has existed in one form or another for the last 150 years. Initially charged with collecting only cross-sectional data, in 1972 NCES launched their first nationally representative longitudinal survey, the NLS:72, with the goal of providing actionable data to both policymakers and researchers regarding the educational, work, family, and community activities of the high school class of 1972. Since this watershed moment, NCES has continued their longitudinal survey work at not only the secondary, but also the postsecondary level; specifically, with the Beginning Postsecondary Students Longitudinal Study (BPS) and Baccalaureate and Beyond Longitudinal Study (B&B) – both spun off of alternative administrations of the cross-sectional National Postsecondary Student Aid Study (NPSAS).

The overwhelming success of these three postsecondary surveys has resulted in a significant amount of attention being focused on the initial design of NPSAS and the technical decisions used to inform this design – especially since these decisions were made almost 25 years ago. In an effort to revisit the original design decisions, in the fall of 2016, NCES and the National Institute of Statistical Sciences (NISS) convened a small panel of experts to address three specific research questions: (1) Do we need to revisit the design decisions made for NPSAS in the early 1990s? (2) Are we adequately using techniques to reduce error? (3) Are there design changes that NCES can employ to increase the utility of the studies? After several days of discussion, this panel of experts was charged with putting their thoughts into writing and the result is this technical manuscript.

The organization of this report loosely follows the order of the three research questions in that the first section presents a total survey error perspective for redesigning NPSAS, followed by a section on sample design and sample error that discusses estimation goals, the two-stage sample design and composite measures of size, rotation of sample institutions, weights and their design effects, variables used for nonresponse and calibration, and finally, the dual frame sample for veterans. The third section of the paper addresses various types of measurement error and offers recommendations for minimizing this pervasive survey-related problem, while the fourth section discusses issues surrounding weighting and imputation. The penultimate section of the paper shifts gears a bit and discusses the role of administrative records and heterogeneous data sources such as the Department of Education's Central Processing System, the National Student Loan Data Systems, the National Student Clearinghouse, and ACT and SAT test score data. The paper then concludes with a summary of the panel's deliberations and a compilation of the findings that appear at the end of each individual section.

I. A PERSPECTIVE AND A FRAMEWORK

A Total Survey Error Perspective for Redesigning NPSAS

Redesigning a survey presents a rare opportunity to consider all aspects of the survey from the perspectives of both the producer and user. This holistic view of survey quality is sometimes referred to as the *total survey quality perspective*. Total survey quality is comprised of essentially five dimensions – Accuracy, Relevance/Contents, Timeliness/Punctuality, Comparability/Coherence, and Accessibility/Clarity. Much of the discussion of the NPSAS redesign so far has focused on the accuracy of NPSAS estimates which is the responsibility and primary concern of the producer (NCES). However, the remaining dimensions are also important, particularly to users of the NPSAS data. Table 1 provides some of the components that comprise the four user dimensions of quality.

Table 1

User Dimensions and their Components

Relevance/Contents • Outputs (including microdata and other products) • Inputs (content, scope, classifications, etc.)	**Timeliness/Punctuality** • Timeliness of release of main aggregates • Timeliness of release of detailed outputs (including microdata) • Punctuality of data releases
Accessibility/Clarity • Level and timeliness of user support • Ease of data access (including microdata where relevant) • Documentation (including metadata) • Availability of quality reports	**Comparability/Coherence** • Comparability across geography, populations, and other relevant domains • Comparability across time (including impacts of redesign) • Coherence with other relevant statistics (including use of standard classifications, frameworks, etc.)

The dimensions in this table are important considerations in the NPSAS redesign and should be afforded all the attention they merit. However, the focus of this report is the Accuracy dimension. One could argue that Accuracy is key because all other quality dimensions are built upon a foundation of high quality data. Indeed, accuracy or data quality is often taken as a given by data users. For that reason, it should be of primary concern to data producers.

To facilitate the data quality considerations for the NPSAS redesign, issues are presented from a total survey error perspective. This approach decomposes the total error in an estimate (typically a mean or total) into mutually exclusive and exhaustive error components that can be directly associated with survey activities that could potentially generate error in the estimates. The objectives of redesign can then be aligned with these error components to reduce their effects on the total error. Thus, let $\hat{Y}$ denote a survey estimate that is subject to errors from a number of sources. Hypothetically, there exists an "error-free" (unobservable) version of this estimate denoted by *Y*. If the processes producing $\hat{Y}$ were error free and ignoring possible sampling errors, the estimate and the error-free parameter, *Y*, would

agree. The difference between the two can then be attributed to errors in the processes that produce $\hat{Y}$, i.e., the *total survey error*. The total survey error (TSE) includes both the *non-sampling* error and sampling error for the estimate. In this report, the TSE is decomposed into seven components: specification error, frame error, nonresponse error, measurement error, data processing error, sampling error and model/estimation error.

Specification error arises when the observed variable, *y*, differs from the desired construct, *x* - i.e., the construct that data analysts and other users prefer. In survey literature, for example Biemer (2011), *x* is often referred to as a *latent* variable representing the true, unobservable variable and *y* is often referred to as an indicator of *x*. As an example, the question "What was your income in the 2015-2016 academic year?" is subject to specification error because the respondent's definition of "income" may not conform to the researcher's definition; i.e., the construct implied by the question (*y*) may differ from the underlying construct required by the researcher (*x*) as a result of inclusion or exclusion of sources of income specified in researcher's definition. Specification error may be defined as the difference between *y* and *x* (see, for example, Biemer and Lyberg (2003)).

Frame error arises in the process of constructing, maintaining, and using the sampling frame(s) for selecting the survey sample. It includes the inclusion of non-population members (*over-coverage*), exclusions of population members (*under-coverage*), and duplication of population members, which is another type of over-coverage error. Frame error also includes errors in the auxiliary variables associated with the frame units (sometimes referred to as *content error*) as well as missing values for these variables[1].

Nonresponse error encompasses both unit and item nonresponse. *Unit nonresponse* occurs when a sampled unit does not respond to any part of a questionnaire. *Item nonresponse* occurs when the questionnaire is only partially completed because an interview was prematurely terminated or some items that should have been answered were skipped or left blank. *Measurement error* includes errors arising from respondents, interviewers, survey questions and factors, which affect survey responses. *Data processing error* includes errors in editing, data entry, coding, computation of weights, and tabulation of the survey data. *Modelling/estimation error* combines the error arising from fitting models for various purposes such as imputation, derivation of new variables, adjusting data values or estimates to conform to benchmarks, and so on.

Then the total survey error in $\hat{Y}$ can be written as

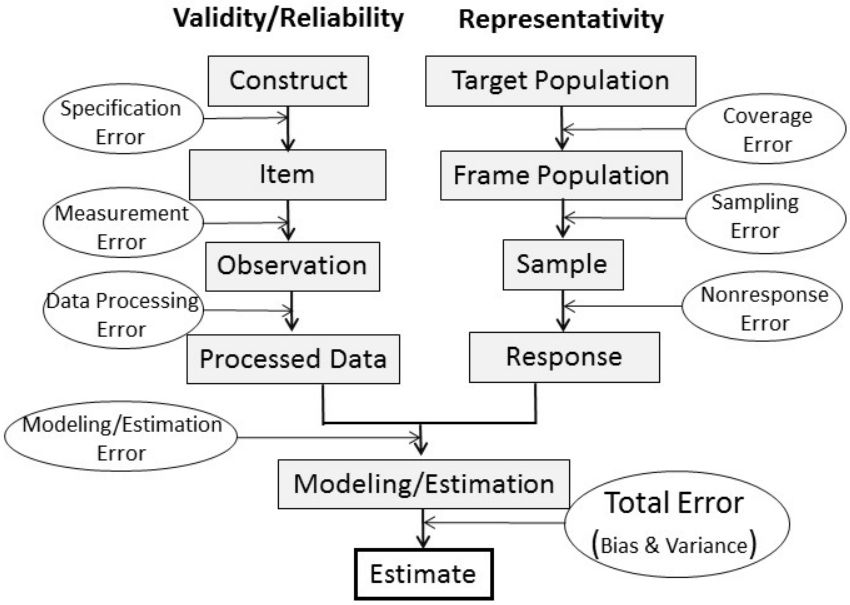
$$\hat{Y} - X = (Y - X) + (\hat{Y} - Y), \text{ or, in words,}$$
$$\text{TSE} = \text{(specification error)} + \text{(other sampling and nonsampling errors)}$$

Under this model, the TSE of an estimate includes specification error as well as the other aforementioned sampling and non-sampling errors. Thus, the specification error in the estimate $\hat{Y}$ is the difference between the expected value of $\hat{Y}$ conditioned on the construct implied by the survey instrument (*Y*) and the population parameter under the preferred construct (*X*). These components can be related to the various stages of the survey design process as shown in Figure 1. This figure illustrates each step in the

[1] In our approach, missing information for frame variables is distinct from missing information for variables collected during a survey. The latter is referred to as survey item nonresponse.

process to develop NPSAS estimates and the type of error that could potentially arise. This figure will be revisited throughout remaining sections as some of the errors associated with primary NPSAS error sources are discussed.

Figure 1 Estimation Process and Components of Total Error*



* Source: Adapted from Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly, 74*(5), 849-879.

Recommendations

1. In redesigning the NPSAS to improve data quality, the NCES should adopt a total error *perspective* that seeks to minimize the error from the primary error sources within specified costs and scheduling constraints. Decision-making in this context should draw on the knowledge and expertise of NCES to identify those sources of error with great impact that are amenable to reduction; a total survey error analysis is not necessarily intended.
2. While the accuracy of the NPSAS data is of primary importance in the redesign, the NCES should strongly consider the so-called user dimensions of total survey quality shown in Figure 1.

II. SAMPLE DESIGN, SAMPLING ERROR AND DESIGN EFFECTS

Overview of Current Design

The current NPSAS design uses two stages with institutions selected at the first stage and students sampled within each sample institution at the second stage. Institutions are sampled with probabilities proportional to a composite measure of size (*pps*) described in more detail below. Students are stratified based on degree program, veteran status, and other characteristics listed in section 2.3. Eleven strata of institutions and 17 strata of students are defined. Consequently, there are potentially 11*17 = 187 strata, although not all institutions will contain all student strata so that the actual number of strata is less than

187. Sampling rates are, apparently, set separately for each of the 17 student strata. General questions that should investigated are:

(a) Whether all of the complexities in the current design usefully contribute to meeting the estimation goals of the survey; and
(b) Whether a simpler design could meet the goals and have benefits like reducing weight variation.

Estimation Goals

NPSAS is intended to produce reliable national estimates of characteristics related to financial aid for postsecondary students, the role of financial aid in how students and their families finance postsecondary education, and the extent to which the financial aid system is meeting the needs of students and families.

Among the general goals of the NPSAS study are to identify institutional, student, and family characteristics related to participation in financial aid programs. Other goals are to analyze special population enrollments in postsecondary education, including students with disabilities, racial and ethnic minorities, students taking remedial/developmental courses, students from families with low incomes, and older students. The distribution of students by major field of study can also be examined. Data can also be generated on factors associated with choice of postsecondary institution, participation in postsecondary vocational education, parental support for postsecondary education, and occupational and educational aspirations.

According to Appendix B in Wine et al. (2013), goals for NPSAS:12 were to "achieve relative standard errors (RSEs) of 10 percent or less and comparable to or less than NPSAS:08 for key national estimates for the full population, undergraduate students, and graduate students and to achieve RSEs of 10 percent or less and comparable to or less than NPSAS:04 for key national estimates for first time beginners (FTBs). A total of 1,104 precision constraints were developed for national-level student estimates." This is a remarkable number of constraints, and it would be interesting to know what percentages of these were actually achieved. Understanding the reasons for this large number of constraints also seems important along with the pros and cons of considerably reducing this number.

Two-stage Sample Design and Composite Measures of Size

Institutions are grouped into 11 strata and a *pps* sample is selected from each. Within each sample institution, students are grouped into as many as 17 strata, depending on the composition of the student body.

Composite measures of size (MOS) allow a *pps* sample of institutions to be selected while obtaining a self-weighting sample of students to be obtained within a set of domains. The NPSAS MOS's account for the desired breakdown of the sample of students in the 17 student strata listed in Table3. If the MOS's are accurate, pre-determined sampling rates of students in each institution will give both a self-weighting sample and desired sample sizes in each domain. Note that this equalizes selection weights only in each student domain and not the final weights, which may differ because of nonresponse adjustment and calibration.

For NPSAS:12 (2011-12), there was a lag between the time period used for MOS calculation, which was 2009 IPEDS Fall and 2008 IPEDS 12-Month Enrollment data, and the time of data collection. Thus, there is some chance of having inaccurate MOS's for some institutions.

Consequently, the effectiveness of the MOS's should be evaluated. Among the aspects that can be investigated are:

Accuracy. Did the MOS's and the student sampling rates determined from them yield the target numbers of sample students? How often did the actual sampling rates have to be adjusted to obtain desired sample sizes?

Edits and imputations. In some cases, institutional data were missing or suspect and were, therefore, imputed or edited. How well did the edited or imputed values reflect what was found in the field for sample institutions? For computing MOS's for the institutional sampling, how well did the imputation of institutional data work and are there better alternatives?

Modifications. In how many institutions are the student sampling rates modified in order to stay within the minimum and maximum workload constraints?

Stratification. Eleven institutional mutually exclusive and exhaustive strata are used and are shown in Table 2. These are based on a combination of control (public, private non-profit, and private for profit), highest degree awarded, and length of matriculation. Seventeen student strata are formed within an institution, although each of these may not be present in all institutions. The 17 are listed in Table 3.

Table 2

Institutional Strata

Stratum	Description
1	Public less than 2-year
2	Public 2-year
3	Public 4-year non-doctorate granting CCBA*
4	Public 4-year non-doctorate granting other
5	Public 4-year doctorate granting
6	Private non-profit, less than 4-year
7	Private non-profit, 4-year, non-doctorate granting
8	Private non-profit, 4-year, doctorate granting
9	For profit, less than 2-year
10	For profit, 2-year
11	For profit, 4-year

* Mainly community colleges that offer bachelor's degrees in a few fields.

Table 3
Student Strata

Stratum	Description
1	Baccalaureate recipients (veterans)
2	Baccalaureate recipients (STEM programs)
3	Baccalaureate recipients (teacher education programs)
4	Baccalaureate recipients (business programs)
5	Baccalaureate recipients (other programs)
6	Other undergraduate students (veterans)
7	Other undergraduate students
8	Graduate students (veterans)
9	First-time graduate students
10	Master's degree students (STEM programs)
11	Master's degree students (education and business programs)
12	Master's degree students (other programs)
13	Doctoral research/scholarship/other students (STEM programs)
14	Doctoral research/scholarship/other students (education & business programs)
15	Doctoral research/scholarship/other students (other programs)
16	Doctoral professional practice students
17	Other graduate students

Evaluating the allocation to strata. Ideally, the allocation to strata should be determined by considering the estimation goals for a critical set of estimates along with the contributions of institution and student sampling to the variances of estimates in that set. A thorough study of the contribution of the stages of sampling to precision should include:

i. Derivation of a theoretical approximation to the variance of estimators of a total, mean, proportion, or other form of estimator that is important to the survey. This generally will involve an assumption that the pps sample of institutions is selected with replacement in order to simplify variance formulas. (Ad hoc fpc's can be inserted to account for substantial sampling fractions.)

ii. Estimation of variance components for each key statistic.

iii. Formulation of an approximate cost function accounting for the cost of sampling institutions and students, including anticipated follow-up cost per complete case (institution and student).

iv. Determination of the optimal allocation for each statistic or for some combination of, say, the relvariances for a set of important estimates.

v. Comparison of the optimal allocation(s) to the one currently being used.

Based on Appendix B in Wine et al. (2013), this sort of study was done for NPSAS:12. Both a cost function and a relvariance model were developed by RTI as part of designing the sample. As a result, the future work might be geared toward evaluating whether the estimation goals for NPSAS:12 were actually achieved given those design decisions and attempting to determine where improvements can be made. This does assume that the same goals pertain in future editions of NPSAS. If the goals change, then

evaluating the optimality of the previous sample will not be that informative. However, the variance components can be used in designing a survey with a revised set of goals.

Rotation of Sample Institutions

Setting up a rotation plan for some strata of institutions could have several potential advantages:

i. Controlling the amount of overlap in the institutional samples between years may reduce the variance of differences over time or of other measures of change; the importance of estimates of change relative to cross-sectional estimates would be determined by NCES;
ii. For non-certainty institutions, rotation would allow a clear promise to be made for how long each would remain in the sample;
iii. Certainty institutions could be informed that they will always be sampled (assuming their relative size remains large). This might encourage the largest institutions to set up data processing systems that would facilitate extraction of data that NPSAS requires.

Rotation may not be appropriate or efficient in the strata of for-profit institutions because this part of the universe is volatile with some types of institutions rapidly going in or out of business. Consequently, the for-profit design could remain as it is currently.

Rotation is most straightforward with a stratified sample in which equal probability samples are selected in each stratum, although rotation can also be used for *pps* samples (Ohlsson, 1992; Ohlsson, 1995). As noted above, NPSAS does use a composite measure of size for sampling. If, in some strata, this MOS was not particularly effective or if strata based on MOS could be created that have narrow ranges of the MOS, then stratified, equal probability sampling might be an option.

Rotation does have some disadvantages. It introduces additional complications to the current sample selection plan. Plus, it would require supplementation for new institutions in the strata that rotate. The student composition in an institution will also change over time, as would the composite MOS's. Thus, the efficiency of size-based strata would probably decrease over time.

If overlap in the non-certainty part of the sample were desirable, an alternative would be to select using a *pps* method that maximizes overlap with the previous sample when a new cross-sectional sample is selected. Several options are available for doing this with a relatively simple one being Kish and Scott (1971).

Weights - Nonresponse Adjustment and Calibration

General form of estimators. Since there is a large amount of auxiliary information known about both institutions and students, NPSAS can make good use of model assisted estimation. NPSAS now does this via the WTADJUST procedure in SUDAAN. This procedure can adjust for nonresponse and calibrate to population totals of auxiliary variables. This type of model assisted estimator can both reduce variance and adjust for nonresponse bias (Särndal et al., 1992).

By picking proxies on the frame for key outcomes and using the proxies as estimation targets, different weighting schemes and selections of auxiliary variables can be evaluated. This can be done for institution-level data but possibly not for student-level estimates unless individual student data are available from some administrative source. The evaluation can be done by comparing the full sample estimates and the

nonresponse adjusted estimates using the frame variables to calculate nonresponse bias. In addition, variances can be calculated to understand the effect of the adjustments on variances of estimates. Together the variance and the bias can be used to calculate a mean squared error. These three measures—bias, variance, and mean squared error—can be used to evaluate effectiveness of the weighting schemes.

Steps in Weighting. NPSAS uses 11 weighting steps that account for sampling, nonresponse, and other complications that often arise in survey estimation. Wine et al. (2013) thoroughly document the steps used in NPSAS:12. The particular steps were:

1) Institution base weights
2) Institution multiplicity adjustment (designed to account for institutions that had more than one chance of selection due to mergers that occurred after the sampling frame was constructed)
    a. In 2011-12 only 12 institutions received this adjustment.
3) Institution nonresponse adjustment
4) Institution post-stratification adjustment
5) Student base weight
6) Student multiplicity adjustment (designed to account for students who attended more than one NPSAS institution during a year and, as a result, had more than one chance of selection)
7) Student unknown eligibility adjustment
8) Student not located adjustment
9) Student refusal adjustment
10) Student other nonresponse adjustment
11) Student post-stratification adjustment

Each step is motivated by standard concerns in design-based sampling, although evaluating the effects of some of the steps would be worthwhile. A proliferation of steps will tend to increase the variation in weights – perhaps unnecessarily. Simplification of some of the steps may be possible, or, at least, evaluation can be done to determine whether all complexities are really improving the bias or precision of estimates.

Institution nonresponse. One step that might be a candidate for simplification is the nonresponse adjustment for institutions. Table 48 of Wine et al. (2013) list the following variables that were used for institutional nonresponse adjustment in NPSAS:12:

1) Institution type (10 levels)
2) Carnegie classification code (6 levels)
3) Institution region (8 levels)
4) Percent receiving federal grant aid (5 levels)
5) Percent receiving state/local grant aid (4 levels)
6) Percent receiving institution grant aid (4 level)
7) Percent receiving student loan aid (4 levels)
8) Percent enrolled: Hispanic (4 levels)

9) Percent enrolled: Asian or Pacific Islander (4 levels)
10) Percent enrolled: Black, non-Hispanic (4 levels)
11) Total undergraduate enrollment (5 levels)
12) Total male undergraduate enrollment (4 levels)
13) Total female undergraduate enrollment (4 levels)
14) Total male graduate enrollment (4 levels)
15) Total graduate enrollment (5 levels)
16) Total male graduate enrollment (4 levels)
17) Total female graduate enrollment (4 levels)
18) Average net price among students receiving grant or scholarship aid (5 levels)
19) Degree of urbanization (10 levels)
20) Historically Black college or university (2 levels)
21) Hispanic-Serving Institution (2 levels)
22) CHAID segments based on region, Carnegie classification, and institution type (7 levels)

The unweighted institution response rate was 88% in 2011-12. There were 210 non-respondents, i.e., institutions that did not provide enrollment lists. Thus, using 22 variables with a total of 109 levels may be more than necessary to adequately adjust for institutional nonresponse. Use of a large number of variables and categories could also result in excessive weight variation without an accompanying reduction of bias and variance.

Student nonresponse adjustments. Student nonresponse was adjusted for in three stages—inability to locate the student, interview refusal, and other nonresponse. The reasoning given for separating these was that response propensity could depend on different factors that should be accounted for in adjustment. This leads to many different covariates being used for adjustment. The not-located adjustment used 22 variables; individual adjustments ranged from 0.72 to 19.24 (Wine et al., 2013). The refusal adjustment used 23 variables and ranged from 1.00 to 3.78. The other-nonresponse adjustment used 25 variables and ranged from 1.00 to 62.83. The ranges for the first and third adjustments seem especially large. Whether the wide range of adjustments is needed to reduce nonresponse bias should be evaluated. The amount of bias reduction could offset any increase in variance, but this can be part of an evaluation.

Student post-stratification. A final calibration step adjusted student weights based on a list of qualitative and quantitative variables, including types and amounts of loans or grants they received (Stafford, Pell, PLUS) and student level (undergraduate, graduate) by institution type. These adjustments ranged from 0.03 to 73.6. The range here again seems quite large.

Evaluation of Weighting Design Effects

Evaluating the distributions of the weighting factors and weights after each step of the weighting process can provide an understanding of how each step is changing the weights. The wide range of the adjustments detailed above must contribute to increasing the design effects due to weighting. While it is possible that each of the adjustments being used does contribute to either correcting for nonresponse

bias or reducing variances, evaluation should still be done to determine if this is the case. The general steps in the appraisal would be:

i. Compute the design effect due to weighting, deff(w), after each of the 11 weighting steps. That is, compute the quantity, 1 + relvariance of weights, based on the product of weights up to and including the step;

ii. Identify any steps that contribute substantially to increasing deff(w);

iii. Repeat these computations within institution strata and across the full sample;

iv. Compare deff(w)'s to directly computed design effects for key statistics after each step. That is, calculate the variance of an estimate accounting for the complexities of the sample design and weighting method divided by the variance of the estimate that would be obtained from a simple random sample of the same size. To be able to compute the direct deff's after each step, these calculations would probably have to be done using administrative or other data that are available for all institutions and/or students.

The four steps obviously produce many statistics to examine and summary statistics will be necessary. Boxplots of weights after each step should be useful, although decisions will be needed about which combinations of institution and student strata are worth studying.

Although deff(w) can be a useful statistic for evaluating weighting systems, it often does not reflect the real effect on survey estimates of weighting steps. Step (iv) above is intended to partially address this issue.

More importantly, an attempt should be made to devise a measure of mean squared error to use in the evaluation of the weighting steps. If a weighting step contributes unreasonably to increasing variances and/or does not reduce nonresponse bias, then the details of that step should be examined to determine the source of the problem. For example, are there a few extreme adjustments, or are too many variables or levels of variables being used? These effects may differ depending on the statistic and whether an estimate is for a domain or the full population. Thus, evaluations should be made for a collection of important estimates.

Selection of Variables used in Nonresponse and Other Adjustments

As noted earlier, the nonresponse and, to a less extent, the calibration steps use many variables. The variables probably vary in predictive power, and a smaller set could predict the probability of responding for institutions or students as well as does a more elaborate model. Likewise, a smaller set of calibration variables could predict analytic variables as well as the larger set now used. Examining whether the nonresponse and calibration models can be reduced seems worthwhile. A side effect of reducing the models could be elimination of convergence problems in weight calculation. This may be especially true for replicate weights.

Overlap in class variables used for nonresponse adjustment. Investigating the overlap in the class variables in the weighting steps to adjust different types of nonresponse such as unknown eligibility, refusal, etc. and in post stratification could suggest that there is not a need for splitting the weighting into different steps. This could lead to fewer weighting adjustments, which, in turn, may lead to reducing unnecessary variation in weights.

Relation of class variables to response and key outcomes. Little and Vartivarian (2003, 2005) show that class variables used in nonresponse adjustments need to be correlated with both response and key outcomes to be effective in the reduction of nonresponse bias. In light of this, the relationship between the variables that are used to create nonresponse adjustment cells need to be evaluated to ensure that the variables are related to both response and key outcomes. This can be done by fitting regressions to evaluate (i) how well the variables predict survey outcomes or proxies of outcomes found on the frame and (ii) how well the variables predict whether institutions and students respond.

Alternatives for estimating propensities of response. Interactions to be used in the models are now identified using CHAID, which is generally considered to be outmoded and inferior to more modern methods like classification and regression trees (CART). In the case of the nonresponse adjustments, all that is necessary is a predicted probability of response which can be obtained from methods that have been shown to be more robust than CART like random forests, cforest, or bagging (James et al. 2014; Strobl et al. 2007). These alternatives should be worth evaluating to determine whether they predict the probabilities of response better than CHAID. The predicted probabilities of response can be sorted, grouped into a limited number of classes, and a single adjustment value used in each class. This is a standard method of reducing the variation in weight adjustments (Little, 1986; Valliant et al., 2013). How this method compares to SUDAAN's WTADJUST should be evaluated.

Dual Frame Sample for Veterans

In the current design veterans receive special attention by being assigned to special strata (see Table 3). If institution records do not accurately identify veteran status, then the target sample sizes may not be achieved and there is a risk of under-coverage. A dual frame design might be considered if the estimates for veterans were not as precise as desired. The VA list of veterans receiving financial aid in the first-stage sample of institutions would serve as the main frame for the veterans' sample. Other veterans not getting aid would be picked up in the regular sample. Among the advantages of using a special frame for veterans are:

- Better control of veterans' sample size, especially of those receiving financial aid;
- Institutional student lists do not always identify whether students are veterans, which creates a stratum-jumper problem, i.e., veterans can be sampled as part of non-veteran strata. A special sample of veterans would reduce this problem;
- If veterans cease to be a group of special interest, then the VA sample can be dropped.

Disadvantages of a dual frame approach are that:

- It adds complications to estimation, particularly variance estimation;
- Variation in weights may increase if different sampling rates on the different frames are used;
- It is limited by quality of the VA list;
- Contact information on the VA list may be out-of-date and would have to be updated for each veteran;
- Current institution attended is required if the veterans' sample is restricted to institutions in main sample. This may not be up-to-date on the VA list.

Recommendations

1. Evaluate precision goals. Are all 1100+ precision targets necessary to meet the major goals of the survey and could reduction of the number of targets lead to important cost savings? A related topic would be to determine whether all goals set for NPSAS:12 and NPSAS:16 were actually met.
2. Evaluate allocation to institutional and student strata. By estimating variance components and updating the NPSAS cost model, it can be determined whether the most efficient allocation of institutions and students to strata is being used.
3. Explore rotation of sample institutions. Rotation of non-certainty, not-for-profit institutions could increase overlap between surveys and reduce variances of estimates of change. In addition, it could be used to control burdens on some institutions
4. Evaluate composite measures of size. The composite measures of size are somewhat out-of-date by the time student sampling at institutions occurs. In some cases student data at the institutional level are missing and must be imputed. How well the composite MOS's are related to actual student counts and how often within-institution sampling rates have to be adjusted to account for inaccurate MOS's can be investigated.
5. Evaluate weighting steps. Design effects due to weighting can be computed following each of the 11 weighting steps to determine if any steps add unnecessarily to weight variation. Using frame or other data available for both respondents and non-respondents, a study should be devised to investigate the contributions of each step to bias reduction, variance increase or decrease, and the contribution to mean squared error for a set of important estimates. The selection of covariates selected for use in nonresponse adjustment and calibration should also be validated. For example, modern regression tree analysis can be used to determine the predictive power of these covariates both singly and severally.
6. Explore whether a dual frame design with a special frame for veterans is worthwhile. A dual frame design would entail using a VA list of veterans that receive education benefits to select the bulk of the veterans sample and picking up other veterans from the regular student sample. This could control the veterans' sample size better and reduce some under-coverage. Whether the added complexity would be worthwhile would have to be determined by NCES.

III. MEASUREMENT ERROR

Overview

The section briefly overviews a few key concepts behind measurement error, briefly reviews the very extensive measurement error work undertaken in 1997, and provides suggestions for future efforts to quantify and ameliorate measure error in NCES surveys.

Validity. Broadly speaking, validity is a measure of how close a respondent's answer Y_i is to a true value μ_i. Validity can refer to a situation where μ_i is a latent construct, such as overall health, being measured by one or more specific questions about perceived health or health conditions, or where μ_i is a true

value possibly reported with error, such as reported total income on an income tax form. In the former setting, multiple items indexed by α are modeled as:

$$Y_{\alpha i} = \lambda_\alpha \mu_i + \varepsilon_{\alpha i}$$

where λ_α is a measure of the validity of Y_α , with values close to 1 indicating high validity (Cronbach and Meehl 1955). Factor analysis or structural equation models can be used to estimate λ ; in the case of two measures, under constraining assumptions that $\lambda_1 = \lambda_2$ and $\text{var}(\varepsilon_{1i}) = \text{var}(\varepsilon_{2i})$, a validity measure is given by

$$\rho_{Y_1 Y_2} = \frac{\sigma_{Y_1 Y_2}}{\sqrt{\sigma_{Y_1}^2 \sigma_{Y_2}^2}}$$

In the latter setting, the response process is conceptualized as being repeated many times, with the data given by Y_{it} , the observed data for the *t*th "trial" of a survey, and

$$Y_{it} = \mu_i + \varepsilon_{it}$$

where ε_{it} is the deviation between the true value and the response at a given trial. The validity of the response Y is given by the correlation between Y and μ :

$$Validity(Y) = \rho_{Y\mu} = \frac{\sigma_{Y\mu}}{\sqrt{\sigma_Y^2 \sigma_\mu^2}}$$

Values close to 1 indicate measures with high validity.

For estimation, the population covariances and variances are replaced with their sample estimates. Note that the validity estimate assuming a true value requires these true values to be available from an outside source, whereas the latent model is estimated only from the data.

Non-response bias. In the construction above, we assumed the mean of the error terms was 0. If this is not the case, systematic bias will occur:

$$Bias(\bar{Y}) = \frac{1}{T}\sum_t \frac{1}{N}\left[\sum_i \left(Y_{it} - \mu_i\right)\right]$$

In a single study, where true values are available, the estimated bias is simply the measure of the differences between the observed and true values.

Reliability. Reliability focuses on whether respondents are stable in their answers. Hence the key estimate of reliability is how large the variability is in the true responses varies relative to the variability in the "trial-level" errors

$$reliability(Y) = \frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma_\varepsilon^2}$$

where the expectation for σ_ε^2 is taken with respect to both the trial and the population; a common alternative measure is the index of inconsistency:

$$IOI = 1 - reliability(Y) = \frac{\sigma_\varepsilon^2}{\sigma_\mu^2 + \sigma_\varepsilon^2}$$

(Groves et al., 2009). In practice, reliability estimation usually requires some form of a reinterview, with the timing of the interview spaced to allow independence in the trial errors but not so long as to change the underlying true value. Under these assumptions, an estimate of σ_ε^2 can be obtained as

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{2n}\sum_i (y_{i1} - y_{i2})^2$$

(the gross difference rate or GDR), and an estimate of IOI obtained as

$$\frac{GDR}{2s^2}, \text{ where } s^2 = \frac{1}{2}\left(s_1^2 + s_2^2\right),$$

where s_t^2 is the variance estimator of *Y* for the t^{th} interview.

Interviewer effects. It is well-known that estimates of key population parameters tend to vary between interviewers (Hanson & Marks, 1958; Rice, 1929). This between-interviewer variance affects survey estimates in a manner similar to the design effects introduced by cluster sampling, that is the multiplicative increase in the total variance of an estimated mean can be estimated as *deff* = 1 + ρ_{int}(*m* - 1), where *m* is the average number of interviews conducted by individual interviewers and ρ_{int} is the within-interviewer correlation in answers elicited to a particular survey question (O'Muircheartaigh and Campanelli 1998). In multistage designs where interviewers are clustered within the highest-level PSU, this interviewer effect is automatically incorporated into the design effect; otherwise it must be incorporated by treating the interviewer as a unit of clustering. Standard ANOVA or random effects models can be used to estimate interviewer level variance when interviewers are randomized to subjects (interpenetration). The priority for studying interviewer effect in NPSAS depends on the number of interviews conducted by phone, the potential effect size and consequently the relative contribution to total survey error.

Mode effects. In multi-mode studies, respondents may have consistent differences in their responses under differing modes (Hochstim, 1967; Tourangeau & Smith, 1996). For example, respondents to a mail or self-completed portion of a CAPI method might answer differently to a telephone or interviewer conducted CAPI method, especially if the questions involved are of a sensitive nature. Assessing mode effects is best done under randomization: subjects are randomized to one mode or the other, and standard methods for analyzing randomized trials can be used to assess the impact of mode (Kreuter et al. 2008).

Adjustments for mode effects are less common and are the subject of recent research: methods that use missing data methods in the context of causal inference (creating "potential outcomes" under each mode and treating the mode not received as missing) show promise (Kolenikov & Kennedy, 2014), although the choice of a "gold standard" remains an issue unless true values are available from an outside data sources.

Review of Previous 1997 NCES Work

NCES completed an extraordinarily thorough review of measurement error in their many surveys, focusing on measures of validity, non-response bias, and mode effects, using record check and multiple indicator studies to assess validity and response bias, reinterview studies to assess validity and reliability, and cognitive studies to do more qualitative assessments of sources of measurement error (National Center for Educational Statistics, 1997).

Reinterview studies. Reinterview studies were the most common studies undertaken by the NCES in the 1997 report. Twenty-seven NCES reinterview studies were conducted, including studies for the NHES (5), RCG (1), SASS (13), R&B (1), BPS (3), NPAS (3), and NSOPF(1), with resample percentages ranging from 1-49% (most were in the 5-15% range). Reinterviews were conducted for both field tests and full scale studies, with time lags of a few days to a few months. Modes were largely identical, although some studies switched modes, possibly confounding mode effects with other forms of measurement error. Reinterview studies generally focused on measures of reliability, reporting GDR and IOI measures, using a cutoff of 0.2 for high reliability and 0.5 for low reliability for the latter. A full review of this work is beyond the scope of this summary, but these studies identified key questions that appeared to have issues with reliability, which, depending on the survey, ranged from 2 or 3 to over a dozen items. Summary measures for GDR and IOI were usually reported, although occasionally reliability was stratified in some fashion (for example, degree earned by the teacher in SASS). (Note that the 1997 report has a chapter on response variance and validity reliability. The results presented in both can be interpreted as being a form of reliability analysis.)

Record check studies. RCG:91 and SASS Teacher Transcript studies were subjected to record check studies to assess validity and response bias. These were "forward" record checks, where external data sources (e.g., certification records) were obtained only for respondents in the survey.

Multiple indicator studies. Multiple indicator studies were conducted for HS&B, NELS:88 (base year), and NSOPF-93. These studies were used to assess validity as well as response bias. For validity, the focus of these studies was on the comparison of student and parent responses or twin responses under the assumption of equal validity and variance, so that the correlation $\rho_{Y_1Y_2}$ was used to assess validity, with values below 0.5 to indicate poor validity and 0.8 and above good validity. Again, the focus was on reporting items with low correlation/validity, which was about 15% of the items in the survey. Analyses stratifying by respondent type (for example, there was higher validity for seniors and parents than sophomores and parents) were also conducted. Validity measures of student reporting of grades and courses taken were also conducted, comparing with transcripts. Response bias on child reporting of parents' income and occupation (treating parent responses as truth) were also conducted.

Cognitive studies. Finally, the 1997 report detailed several cognitive research results using the BPS, NHES, and SASS. These studies involved careful probing of respondents to determine how respondents interpreted question and self-reported difficulties with memory retrieval and question interpretation. Reviews were also conducted of interviewer behavior, noting wording changes, question clarifications, and displays of affect.

Recommendations

1. Repeat of 1997 measurement error report. The 1997 report is an extremely comprehensive and thorough assessment of measure error. This report can serve as an excellent template for another, contemporary, thorough review of measurement error in the modern NCES system.
2. New work. Two areas that were not covered in the 1997 report are interviewer effects and mode effects. While interviewer effects were a recognized concern long before 1997, the 1990s and 2000s were something of a "Dark Age" for research and assessment in this area; research on mode effects was largely in its infancy then, since multi-mode studies were just starting to become common in response to declining response rates and new technology development.

Both of these areas should be assessed directly in future research. While measures of reliability indirectly include assessment of interviewer effects if interviewers are switched in follow-up studies, a more direct assessment should be made of interviewer effects using analytic methods that account for clustering. (Ideally interviewer assignment should be randomized to the degree possible to avoid confounding of respondent assignment and non-response with measurement error induced by interviewers.) In addition, for studies in which interviewers are not nested within PSUs, interviewer IDs should be included in analytic datasets, and user documentation should emphasize the need to account for the measurement error introduced by interviewer effects (Elliott & West, 2015). For mode effect, particular web versus telephone, randomized assignment of modes should be used with a random subset of the sample to assess potential mode effects, with record check studies used where possible to determine which mode is closer to a "gold standard". If randomized studies are not possible, consideration should be given to analyses that treat answers under each mode as a "potential outcome," using missing data methodology to estimate the counterfactual results under a study that would be consistently under one mode or the other, to assess mode effects (Kolenikov & Kennedy, 2014).

Finally, while NCES has an excellent assessment of measurement error, it is perhaps less clear how this assessment is being used to improve questions and reduce measurement error going forward. The 1997 report did attempt to summarize their findings, noting that items with poor recall, items asking for specific dollar amounts, items with large numbers of response categories, and attitudinal items tended to have lower reliability. Several questions were also successfully revised to increase reliability. Nonetheless, a more routine system to review, revise, and perhaps even eliminate questions that are subject to large amounts of measurement error might be considered for the future.

IV. IMPUTATION AND ADMINISTRATIVE DATA

The use of administrative data and imputation are closely linked in the NPSAS. Both are used to address missing data at the item level. In addition, administrative data are used to augment the survey data and a NPSAS based solely based on administrative records has been contemplated. This makes it critical that both the imputation methods and the use of administrative data critical to the quality of the NPSAS should be evaluated going forward.

Imputation

According to the 2011-12 NPSAS Data File Documentation a weighted sequential hot deck was used to impute for missing data. Class variables were selected using classification trees. The documentation of the imputation method is good in the Data File Documentation. One additional thing that might be included to help data users better understand how imputation effects their analysis is the list of class covariates used for each variable and the order in which the variables were imputed.

In general the weighted sequential hot deck seems to be a reasonable imputation method for variables with small rates of missing data (<10%). If used for variables with more than 10% missing and the imputed values are treated as true responses, variances can be biased downwards even if the imputation model is correctly specified. There are a few ways that this can be addressed such as fractional imputation and multiple imputation. The reported percent missing data rates in Appendix M of the Data File Documentation show that a majority of the of the variables have imputation rates greater that 10%, and a large number have rates over 30%. Because of this the current method of ignoring the effect on variances is likely to be biased.

Although it is not the only way of dealing with the downward bias in variance estimation, multiple imputation (MI) is a popular way of dealing with this problem. In general terms, multiple imputation is implemented by creating multiple datasets repeating the imputation for each dataset. From these datasets two types of variances can be calculated: the variance within each dataset and the variance between the datasets for a given estimate. Once each of these variances is calculated the following formula can be used to calculate the total variance (*T*):

$$T = \bar{W} + \left(1 - \frac{1}{m}\right) B$$

where $\bar{W}$ is the average of the variance within each of the datasets, *m* is the number of datasets that are generated and *B* is the between data set variance (Rubin, 1987).

One of the challenges to moving to multiple imputation is the need to change the imputation procedures. It should be noted that taking multiple draws out of a hot deck does not adequately capture the between variance (Andridge & Little, 2010). One way that this can be done is by using a generalization of the hot deck imputation methods already being employed using an Approximate Bayesian Bootstrap. This can be computationally heavy but with modern computing power this should be manageable. Other federal agencies with similarly extensive survey data bases (e.g., NASS) use MI and might serve as models for conquering the challenges of data complexity and cost cited in previous decisions for NPSAS. NCES does use MI in some of its other surveys - NAEP, in particular - so that a move to MI would not be unprecedented within the agency.

Administrative Records and Heterogeneous Data Sources

Administrative Data. In addition to survey responses from institutions and students, student data for the NPSAS:12 were obtained from administrative databases: US Department of Education Central Processing System (CPS), the National Student Loan Data Systems (NSLDS), the National Student Clearinghouse (NSC), ACT test score data, and SAT Reasoning Test score data. The administrative data provided student-level information that was used to supplement the interview and/or to reduce response burden, provided

values for some item-level data that were missing either from an institution or a student response, and served as an accuracy check of similar information from other sources. However, the administrative data is incomplete as detailed in the sections below.

U.S. Department of Education Central Processing System (CPS). To obtain federal, and often state and college, financial aid, a student must complete a Free Application for Federal Student Aid (FAFSA), which gathers information on the financial status of the student and his/her family. The FAFSA data, which are stored in the CPS, are collected for the NPSAS after the student sample is selected but before data collection begins. For NPSAS:12, the student sample was matched against the CPS data for the 2011-2012 financial aid year using the CPS ID, which is a student's social security number concatenated with the first two letters of the student's last name. If the social security number was missing, the student record was not matched to the CPS.

Overall, 77.4% of the students in the NPSAS:12 sample, had submitted an application for student aid (FAFSA) and consequently matched a CPS record. The proportion of students with CPS record (i.e., FAFSA filed) varied by type of student and institution. Among all undergraduates, this rate was 81.5%, but it was 88.1% for potential first-time beginning undergraduates. In contrast, the proportion of graduate or first-professional students in the sample with FAFSA filings in the CPS database was 52.0%. For four-year, doctorate-granting private non-profit institutions, 67.2% of the student sample associated with those institutions had CPS records; however, this rate was 92.1% for the student sample members from private for-profit two-year colleges (IES 2013, Table 43, p. 85).

National Student Loan Data Systems (NSLDS). The NSLDS is the Department of Education's central database for student aid, comprising all students who received a loan. For NPSAS:12, student-level data on the nature and amount of Pell Grants and on Federal student loans received were collected. Data were collected twice from the NSLDS for NPSAS:12, first for preliminary data and then after data collection was complete in an effort to have the most current data available. The NPSAS:12 sample was matched to the NSLDS using name, SSNs, and date of birth. Only students who had a valid grant or loan record in the NSLDS could be matched.

For the full NPSAS:12 sample, 59.4% had (matched) NSLDS records. Among all undergraduates, the match rate was 64.5%, and it was 30.7% for graduate or first-professional students. The match rates were 53.8%. 37.2%, and 79.4% for public, private nonprofit, and private for-profit institutions, respectively. For sample members associated with four-year, doctorate-granting private non-profit institutions, 27.7% of the student sample matched. In contrast, the match rate was 85.6% for the student sample members from private for-profit two-year colleges (IES 2013, Table 44, p. 86).

National Student Loan Clearinghouse (NSC). Data on institutions attended, enrollment dates, and degree completions for the student sample were obtained from the NSC Student Tracker service for students in the NPSAS:12 sample. Data were collected from the NSC once, toward the end of the data collection period. Student name, SSN, and date of birth were the matching variables. Only those students whose institution was a participant in the NSC could be matched, consequently failure to identify an NCS record cannot be taken to mean missing at random.

The match rate for all students in the NPSAS:12 sample to the NSC database was 79.4%. It was 78.0% among all undergraduates; for graduate or first-professional students, the match rate was 86.4%. The match rates for public and private nonprofit institutions were similar at 87.9 and 84.7%, respectively. For

private for-profit institutions, the match rate was lower at 62.9%. For sample members associated with public, less-than-two-year institutions, the match rate was 29.8%; however, it was 91.0% for four-year, non-doctorate-granting public institutions (IES 2013, Table 45, p. 87).

ACT and SAT test score data. For each student in the NPSAS:12 sample matching an ACT test record between 2005-06 and 2010-11, the most recent ACT survey data and scores were collected. Data were collected after data collection using name, SSN, date of birth and sex as matching variables. The overall match rate was 21.7%. Among undergraduates, the match rate was 24.6%, but it was only 5.4% graduate or first-professional students. For four-year, doctorate-granting public institutions, 34.4% of the student sample associated with those institutions matched, and it was 10.6% for the student sample members from private for-profit two-year institutions (IES 2013, Table 46, p. 89).

For the SAT Reasoning Test, the most recent student records and questionnaire data were obtained for high school graduation years 2006-2011. Data were collected after data collection using the matching variables of name, date of birth, SSN, and sex. The match rate for the full NPSAS:12 sample was 15.7%. Although the match rate for graduate or first-professional students was 0.6% the undergraduate match rate was 18.4%. For public, private nonprofit, and private for-profit institutions, the match rates were 17.6, 25.8, and 7.7%, respectively. For four-year, non-doctorate-granting public institutions, 28.4% of the student sample associated with those institutions matched, and it was 3.9% for the student sample members from public less-than-two-year institutions (IES 2013, Table 46, p. 89).

When considering matches to ACT, SAT Reasoning Test, or both, the sample match rate was 31.5%. For undergraduates, it was 36.1%, and it was 5.8% for graduate or first-professional students. The combined match rates for public, private nonprofit, and private for-profit institutions were 37.0, 41.9, and 17.3%, respectively. For four-year, doctorate-granting public institutions, the combined match rate was 45.3% for the student sample associated with those institutions, and it was 16.1% for the student sample members from private for-profit, less-than-two-year institutions (IES 2013, Table 46, p. 89).

Quality Assessment of Administrative Data. Although the quality control processes for the student and institutional data collected for NPSAS:12 are described in IES (2013), the methods used for assessing the quality of the administrative data are not known to this panel. The quality of all data, including the administrative data, needs to be assessed and the processes to be documented. For example, what biases, if any, are present in the administrative data? How are data prioritized when data sources provide different information about a student?

The administrative data are used to reduce respondent burden by pre-populating some responses, such as date of birth, into the survey instrument. The extent to which this is done and which variables are considered for pre-population are not specified. Documenting this and the extent to which burden is reduced would be helpful.

Administrative data are used to provide information on non-respondents and in the imputation process. The use of trumping rules, which are priorities for use of administrative data for a particular item, was briefly mentioned during the panel meeting. The exact rules and the processes by which they were determined were not provided. However, a multivariate approach of using all administrative data related to a particular response may be better than trumping rules which were explained as *ad hoc,* using criteria sequentially to identify a single “best” source rather than jointly to take into account multiple complementary or consistent sources. Without actually evaluating their effectiveness, the relative merit

of each criterion or of the sequencing or the frequency of conflicts within the sequence of trumping rules is not known.

Administrative data are increasingly important in all surveys. NCES is fortunate to have five databases from which they can draw data to supplement the survey data collected from students and institutions. Considering how to use these data to improve estimates, increase the utility of the data, reduce respondent burden or using administrative data for imputation is an important activity. The quality of the administrative data and the potential bias their use may introduce should be carefully evaluated.

One way to evaluate these different data sources for use in direct substitution is through the lens of imputation. If one is evaluating whether *X* can be used to directly substitute *Y* and *X* and *Y* are continuous, then substitution can be represented as the following linear model:

$$Y = \beta_0 + \beta_1 X$$

by assuming that $\beta_0 = 0$ and $\beta_1 = 1$. This is a strong assumption. This assumption could be tested using simple regression models fitted with administrative data and reported survey data from respondents and then using a standard test to see if $\beta_0 = 0$ and $\beta_1 = 1$.

Recommendations

1. The imputation documentation for NPSAS should include the list of class covariates used for each variable and the order in which the variables were imputed.
2. Imputation methods should be used that account for variance induced by imputation. There are a few ways that this can be addressed such as fractional imputation and multiple imputation.
3. With the wide use of administrative data in NPSAS the quality of the administrative data should be assessed and documented and variables from the administrative data used to pre-populate questionnaires should be documented. Specifically, this documentation should address what biases, if any, are present.
4. Impacts of any missingness in matches to administrative data and of substitution should be assessed and documented. For example, if there are differences with respect to variables of interest, how do these differences affect imputation and subsequent analyses?
5. The documentation should list the trumping rules and provide an evaluation of their utility.
6. Finally, when direct substitutions are treated as response data there are also variance implications that need to be considered. If the distributions of the administrative data are different from the response data or other administrative data being used for substitution, the variances calculated can be biased. This bias can be in both directions as shown in Beaumont et al. (2011).

V. SUMMARY OF FINDINGS

Each of the previous sections of this report provides details of the panel's findings and the rationale for each issue identified for further consideration by NCES. The degree of immediacy of those recommendations varies, some deserving prompt consideration and others requiring time to address.

Those recommendations are given below in abbreviated form separately by section and approximately ordered by the immediacy of their relevance to implementation.
In addition to the specific findings and recommendations below, the panel notes with concern that NCES faces a structural conundrum in regard to the statistical elements of study or survey design. The technical statistical expertise to continue to assure best statistical practices is not currently present in-house, nor is it represented on current survey oversight committees. While contractors do possess this expertise, there is some possibility for conflict of interest as innovation may or may not be in the contractor's own interest. The panel notes that NCES statistical leadership is well aware of this issue and has encouraged innovation and exploration of new statistical methodology for design, imputation, estimation and all phases of analysis.

Recommendations - I

1. In considering redesign of the NPSAS to improve data quality, frame the design from a total error perspective that seeks to minimize the error from the primary error sources within specified costs and scheduling constraints.
2. In addition to accuracy of the NPSAS data, consider the so-called user dimensions of total survey quality shown in Figure 1.

Recommendations - II

1. Evaluate precision goals. Determine whether all 1100+ precision targets are necessary to meet the major goals of the survey and whether reduction of the number of targets could lead to important cost savings. Ascertain whether all goals were met for NPSAS:12 and NPSAS:16.
2. Evaluate allocation to institutional and student strata. Utilize variance components and update the NPSAS cost model to determine the most efficient allocation of institutions and students to strata.
3. Explore use of a dual frame design with a special frame for veterans (using a VA list of veterans who receive educational benefits) to control the veterans' sample size better and reduce some under-coverage.
4. Explore rotation of sample institutions. Rotation of non-certainty, not-for-profit institutions could increase overlap between surveys and reduce variances of estimates of change. In addition, it could help manage burden on some institutions.
5. Evaluate composite measures of size. Since the composite measures of size are somewhat out-of-date by the time student sampling at institutions occurs, some student data at the institutional level are missing and must be imputed. Evaluate how well the composite MOS's are related to actual student counts and how often within-institution sampling rates have to be adjusted to account for inaccurate MOS's.
6. Evaluate weighting steps. Compute design effects due to weighting following each of the 11 weighting steps to determine which, if any, steps add unnecessarily to weight variation. Devise a study to investigate the contributions of each step to bias reduction, variance increase or decrease, and the contribution to mean squared error for a set of important estimates. Use modern

regression techniques to study the selection of covariates for use in nonresponse adjustment and calibration.

Recommendations - III

1. Repeat the 1997 measurement error report. This comprehensive and thorough 1997 report can serve as an excellent template for a contemporary, similarly thorough review of measurement error in the modern NCES system.
2. Evaluate two additional aspects of measurement quality: interviewer effect and mode effect. These two areas were not covered in the 1997 report and should be assessed directly in future research. Use direct assessment of interviewer effects employing randomization where possible to avoid confounding of respondent assignment with non-response induced by interviewers and/or measurement error introduced by interviewer effects. Assess mode effect, in particular web versus telephone, with record check studies when possible to determine which mode is closer to a "gold standard".
3. Utilize NCES's excellent assessment of measurement error to improve questions and reduce measurement error going forward. In addition to an in-depth review, like the 1997 report, consider instituting a more routine system to review, revise, and perhaps even eliminate questions that are subject to large amounts of measurement error.

Recommendations - IV

1. Implement imputation methods that account for variance induced by imputation, for example fractional imputation or multiple imputation.
2. Make trumping rules explicit and provide an evaluation of their utility.
3. Include in the imputation documentation for NPSAS the list of class covariates used for each variable and the order in which the variables were imputed.
4. Assess and document the quality of the various administrative data in NPSAS, specifically addressing what biases, if any, are present. Document the quality of administrative data used to pre-populate questionnaires.
5. To the extent possible, determine whether the students who do not match to any administrative data are comparable to those who match to administrative data; and evaluate the effect of any differences with respect to variables of interest on imputation and subsequent analyses.
6. Examine the variance implications when direct substitutions are treated as response data as differences in the distributions of the administrative data from those of the response data can result in bias (in either direction) of the calculated variance.

VI. REFERENCES

Andridge, R.H. & Little, R. J. (2010). A Review of Hot Deck Imputation for Survey Nonresponse. International Statistical Review, 78, 1, 40-64.

Beaumont, J.F., Haziza, D. & Bocci, C. (2011). Variance estimation under auxiliary value imputation. Statistica Sinica, 21, 515-538.

Cronbach L., Meehl P. (1955). Construct validity in psychological tests, Psychological Bulletin, 52, 281-302.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. Journal of the American Statistical Association, 87, 376-382.

Elliott, M.R., West, B.T. (2015). Clustering by interviewer: a source of variance that is unaccounted for in single-stage health surveys. American Journal of Epidemiology, 182, 118-126.

Groves, R. M. & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly, 74*(5): 849-879.

Groves, RM., Fowler, F.J., Couper M.P., Lepkowski, J.M., Singer, E., Tourangeau, R. (2009). Survey Methodology, 2nd edition. Wiley: New York.

Hanson R.H., Marks E.L. (1958). Influence of the interviewer on the accuracy of survey results. Journal of the American Statistical Association, 53, 635-655.

Hochstim, J. (1967). A critical comparison of three strategies of collecting data from households. Journal of the American Statistical Association, 62, 976–989.

Institute of Education Sciences (IES). 2013. 2011-2012 National Postsecondary Student Aid Study (NPSAS:12): Data File Documentation. U.S. Department of Education, National Center for Educational Statistics. NCES 2014-182. 181 pp.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2014). *An Introduction to Statistical Learning*. New York: Springer.

Kolenikov, S., Kennedy, C. (2014). Evaluating three approaches to statistically adjust for mode effects. Journal of Survey Statistics and Methodology, 2, 126-158.

Kreuter, F., Presser, S., Tourangeau R. (2008). Social desirability bias in CATI, IVR, and Web surveys: the effects of mode and question sensitivity. Public Opinion Quarterly, 72, 847–865.

Little R.J.A. (1986). Survey nonresponse adjustments for estimates of means. International Statistical Review 54(2):139–157.

Little, R.J. & Vartivarian, S. (2003). On Weighting the Rates in Nonresponse Weights. Statistics in Medicine, 22, 1589-1599.

Little, R.J.A. & Vartivarian, S. (2005). Does Weighting for Nonresponse Increase the Variance of Survey Means? Survey Methodology, 31, 161-168

National Center for Educational Statistics (1997). Measurement error studies at the National Center for Education Statistics, NCES 97-464, Washington, DC.

Ohlsson, E. (1992). *SAMU — The system for coordination of samples from the business register at Statistics Sweden: A methodological summary*. R&D report 1992:18. Stockholm: Statistics Sweden.

Ohlsson, E. (1995). Coordination of samples using permanent random numbers. In *Business Survey Methods*, eds. B.G. Cox, D.A. Binder, D.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott. New York: John Wiley & Sons.

O'Muircheartaigh C., Campanelli P. (1998). The relative impact of interviewer effects and sample design on survey precision. Journal of the Royal Statistical Society, A161, 63-77.

Rice, S.A. (1929). Contagious bias in the interview: a methodological note. American Journal of Sociology, 52, 420-423.

Rubin, D.B. (1987) Multiple Imputation for Nonresponse in Surveys. New York: Wiley & Sons.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). Model Assisted Survey Sampling. New York: Springer-Verlag.

Strobl, C., Boulesteix, A., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics* 8(25). http://www.biomedcentral.com/1471-2105/8/25.

Théberge, A. (2000). Calibration and restricted weights. Survey Methodology, 26, 1, 99-107

Tourangeau, R., Smith T. (1996). Asking sensitive questions: the impact of data-collection mode, question format, and question context. Public Opinion Quarterly, 60, 275–304.

Valliant, R., Dever, J. A., and Kreuter, F. (2013). *Practical Tools for Designing and Weighting Sample Surveys*. New York: Springer.

Wine, Jennifer, Bryan, Michael, Siegel, Peter, and Hunt-White, Tracy (2013). *2011–12 National Postsecondary Student Aid Study (NPSAS:12): Data File Documentation*. December 2013. NCES 2014-182, U.S. Department of Education.

VII. APPENDIX

1. AGENDA
2. EXPERT PANEL BIOSKETCHES

AGENDA

Department of Education

ies National Center for Education Statistics
Institute of Education Sciences

NCES NISS NPSAS Design Panel

August 31 – September 1, 2016 · Washington, DC

AGENDA

Technical Expert Panel for NCES on
Study Design Decisions for Postsecondary Sample Surveys

Day 1 – Wednesday, August 31, 2016

8:30am	Introductions
8:45am	Purpose of Panel
9:00am	Setting the Context of Postsecondary Surveys • NPSAS – National Postsecondary Student Aid Study • BPS – Beginning Postsecondary Student Longitudinal Study • B&B – Baccalaureate and Beyond Longitudinal Study • New Initiatives
10:15am	Break
10:30am	Current Design Decisions and Implications • Overview • Sampling Decisions • Non-response Bias, Weighting and Variance Estimation • Theoretical Frameworks for Instrument Design
Noon	Lunch – Executive Session (including development of questions for clarification or added detail from morning sessions)
1:15pm	Panel's Questions for NCES Arising from Morning Sessions
2:00pm	Technical and Practical Challenges and New Approaches • Adaptive/Responsive Design at NCES
3:15pm	Break
3:30pm	Technical and Practical Challenges and New Approaches • Data Analysis Considerations • Total Survey Error Requirements for Contractors
4:30pm	Executive Session
5:00pm	Adjourn

Day 2 – Thursday, September 1, 2016

8:30am	Executive Session
9:00am	Panel's Questions for NCES Arising from Day 1 Sessions
10:00am	Break
10:15am	Executive Session
11:45am	Discussion with NCES (further clarifications, final questions)
12:45pm	Lunch
1:00pm	Executive Session
2:00pm	Adjourn

EXPERT PANEL BIOSKETCHES

***Paul Biemer**, Ph.D., Texas A&M University*

Title: Distinguished Fellow, RTI and Associate Director of Survey Research and Development, Odum Institute, University of North Carolina at Chapel Hill

Dr. Biemer is Distinguished Fellow in Statistics at RTI and Associate Director of Survey Research and Development in the Odum Institute at UNC. He has over 35 years of experience in survey methodology, complex survey design and data analysis and over 100 publications related to these areas. Prior to joining RTI, Dr. Biemer was Head of the Department of Statistics at New Mexico State University and Assistant Chief of Statistical Research Division at the Census Bureau. An internationally recognized expert and researcher in survey methodology, Dr. Biemer's co-authored book, Introduction to Survey Quality is a widely used course text. He co-invented and co-developed Computer Audio Recorded Interviewing (CARI) and pioneered the field of latent class analysis for survey evaluation. His recent book, Latent Class Analysis of Survey Error, is the first book to describe how latent class analysis can be applied to complex surveys, including panel surveys, to evaluate survey error. His current research has focused on the implications of data errors on 'big data' analytics. Dr. Biemer is a Fellow of the ASA and the AAAS and an Elected Member of the ISI. He holds a number of awards for his contributions to the field of survey methodology and statistics, including the including the Morris Hansen Award and the Roget Herriot Award.

***Michael R. Elliott**, Ph.D., University of Michigan*

Title: Professor of Biostatistics, Research Professor of Survey Methodology, University of Michigan

Michael R. Elliott is Professor of Biostatistics at the University of Michigan School of Public Health and Research Scientist at the Institute for Social Research. He received his Ph.D. in biostatistics in 1999 from the University of Michigan. Prior to joining the University of Michigan in 2005, he held an appointment as an Assistant Professor at the Department of Biostatistics and Epidemiology at the University of Pennsylvania School of Medicine, and prior to that as a Visiting Professor of Biostatistics at the University of Michigan School of Public Health and as a Visiting Research Scientist at the University of Michigan Transportation Research Institute. Dr. Elliott's statistical research interests focus around the broad topic of "missing data," including the design and analysis of sample surveys, casual and counterfactual inference, and latent variable models. He has worked closely with collaborators in injury research, pediatrics, women's health, and the social determinants of physical and mental health. Dr. Elliott serves as an Associate Editor for the Journal of the American Statistical Association.

***Fred Galloway,** Ed.D., Harvard Graduate School of Education*

Title: Professor, University of San Diego

Dr. Fred Galloway is currently a Professor and 2012-13 University Professor in the School of Leadership and Education Sciences at the University of San Diego; he also serves as a senior research associate at both the Center for Education Policy and Law and the Caster Family Center for Nonprofit and Philanthropic Research at the university. He received his bachelor's and master's degrees in economics from the University of California, San Diego, and master's and doctoral degrees from the Harvard Graduate School of Education. Dr. Galloway's research interests include the economics of education, higher education policy, and research design and methodology, and he has published more than 60 journal articles, policy reports, and book chapters – including recent articles in Education Administration Quarterly, the Journal of Business Administration Research, the Journal of Research in Leadership Education, Asia Pacific Education Review, and the Handbook of Research on Online Instruments, Data Collection and Electronic Measurements: Organizational Advancements. He has also provided methodological guidance and direction on 102

completed dissertations (54 as chairperson and 48 as associate member), and has served on numerous technical review and advisory panels, including five technical review panels for the National Center for Education Statistics, the Congressional Budget Office Student Loan Advisory Panel, the General Accounting Office Student Loan Advisory Panel, the Lumina Foundation Technical Advisory Panel, and the Technical Review Group for the U.S. Department of Education Follow-Up Evaluation of the GEAR UP Program. In 1999 Dr. Galloway also testified before the US House of Representatives as a friendly witness regarding the results of the national Direct Loan evaluation. In addition to his love of research, he is a passionate teacher who has been recognized with several faculty of the year awards by undergraduates as well as graduate students in both economics and education.

Benjamin Reist, *M.S., George Washington University*

Title: Assistant Center Chief, Center for Adaptive Design at U.S. Census Bureau

Mr. Reist is the Assistant Center Chief for Research in the Center for Adaptive Design at the U.S. Census Bureau. He has also served as the Survey Director for the National Survey of College Graduates. Mr. Reist has ten years of experience working on demographic surveys. He holds an M.S in Mathematics from the University of Vermont and an M.S. in Statistics from George Washington University. He is a Ph.D. candidate in the Joint Program in Survey Methodology at the University of Maryland.

Richard L. Valliant*, Ph.D., Johns Hopkins University*

Title: Research Professor, University of Michigan & Joint Program for Survey Methodology, University of Maryland

Dr. Richard L. Valliant is a Research Professor at the University of Michigan and the Joint Program for Survey Methodology at the University of Maryland. He has over 40 years of experience in survey sampling, estimation theory, and statistical computing. He was formerly an Associate Director at Westat and a mathematical statistician with the Bureau of Labor Statistics. He has a range of applied experience in survey estimation and sample design on a variety of establishment and household surveys. He is also a Fellow of the American Statistical Association and has been an editor of the Journal of the American Statistical Association, the Journal of Official Statistics, and Survey Methodology.

Linda J. Young*, Ph.D., Oklahoma State University*

Title: Chief Mathematical Statistician & Director of Research and Development, USDA's National Agricultural Statistics Service

Linda J. Young is Chief Mathematical Statistician and Director of Research and Development of USDA's National Agricultural Statistics Service. She oversees efforts to continually improve the methodology underpinning the Agency's collection and dissemination of data on every facet of U.S. agriculture. Prior to joining NASS, Dr. Young served on the faculties of three land grant universities: Oklahoma State University, University of Nebraska, and the University of Florida. She has three books and more than 100 publications in over 50 different journals, constituting a mixture of statistics and subject-matter journals. A major component of her work has been collaborative with researchers in the agricultural, ecological, and environmental sciences. She has been the editor of the Journal of Agricultural, Biological and Environmental Statistics. Dr. Young has served in a broad range of offices within the professional statistical societies, including President of the Eastern North American Region of the International Biometric Society, Vice-President of the American Statistical Association, Chair of the Committee of Presidents of Statistical Societies, and member of the National Institute of Statistical Science's Board of Directors. Dr. Young is a fellow of the American Statistical Association (ASA), a fellow of the American Association for the Advancement of Science (AAAS), and an elected member of the International Statistical Institute (ISI).

Panel convened by National Institute of Statistical Sciences

Nell Sedransk, *Ph.D., Iowa State University*

Title: Director, National Institute of Statistical Sciences; Statistics Professor, North Carolina State University

Dr. Nell Sedransk is the Director of the National Institute of Statistical Sciences and Professor of Statistics at North Carolina State University. She is an Elected Member of the International Statistical Institute, also Elected Fellow of the American Statistical Association. She is coauthor of three technical books; and her research in both statistical theory and application appears in more than 60 scientific papers in refereed journals. The areas of her technical expertise include: design of complex experiments, Bayesian inference, spatial statistics and topological foundations for statistical theory. She has applied her expertise in statistical design and analysis of complex experiments and observational studies to a wide range of applications from physiology and medicine to engineering and sensors to social science applications in multi-observer scoring to ethical designs for clinical trials.