Institute of Education Sciences
National Center for Education Statistics

NATIONAL INSTITUTE OF STATISTICAL SCIENCES
TASK FORCE REPORT

NON-RESPONSE BIAS ANALYSIS

TABLE OF CONTENTS

NATIONAL INSTITUTE OF STATISTICAL SCIENCES

NON-RESPONSE BIAS ANALYSIS

EXECUTIVE SUMMARY

The purpose for convening this task force was to assist National Center for Education Statistics (NCES) in understanding the range of methods used currently for nonresponse bias analyses in its data collections, the criteria by which such methods are selected, and other available techniques for Non-Response Bias Analysis (NRBA), including Bayesian methods. Experiments were also conducted to inform the Task Force in reaching its recommendations. (See Appendices A and B). Ultimately, NCES may revise its statistical standards - specifically, Standard 4.4 (http://nces.ed.gov/statprog/2002/std4_4.asp) in light of recommendations by the Task Force.

SPECIFIC RECOMMENDATIONS

The Task Force recommends that NCES employ the NRBA process described in section 1, and specifically that:

1. A *minimal NRBA report* be required for NCES data collections prior to the main data collection, which compares respondents and nonrespondents in terms of available - frame and other) - variables identified on the basis of domain knowledge or other data collections to be related to responses and outcomes of interest in the data. In general, such comparisons can and should be performed using multivariate methods.
2. From the minimal report, NCES determine whether the response bias is *unimportant or important*.[1]
3. In the case of unimportant nonresponse bias, *base weights of respondents be adjusted*, and the data collection and subsequent reporting proceed.
4. In the case of important response bias, *two mandatory additional steps* be performed: (1) Fine-grained analysis, in the same as in the minimal report, for subgroups of importance; and (2) A benchmarking comparison to similar studies.
5. Also, in the case of important response bias, as many as *four optional additional steps* be performed, chosen from (1) Sensitivity analysis; (2) Level-of-effort analysis; (3) Identification of additional predictors of the responses and outcomes of interest, and further comparison of respondents and nonrespondents using them; (4) Follow-up of nonrespondents.
6. Following completion of the mandatory and optional additional steps, NCES decide whether to "discard" or "employ" the data, and if the data are employed, whether to label nonrespondent values as missing, or to reconstruct these values using multiple imputation.

The Task Force further recommends that:

1. Frame and other information pertaining to nonrespondents not be removed from restricted or publicly released databases.
2. The minimal report and the results of mandatory and optional additional steps be made available by NCES to users of the data, possibly in redacted form.

[1] These terms are used for convenience in this report; NCES may wish to use alternatives.

NATIONAL INSTITUTE OF STATISTICAL SCIENCES TASK FORCE REPORT

PREFACE

The National Institute of Statistical Sciences (NISS) was asked by the National Center for Education Statistics to convene an expert Task Force to examine and evaluate methodology used by NCES and its contractors to perform nonresponse bias analyses. Specifically, the Task Force was charged to:

1. Examine and evaluate methodology used by NCES and its contractors for nonresponse bias analysis (NRBA).
2. Articulate best practices for NRBA in various NCES data collections.

The goal is to help NCES understand the range of methods used currently for nonresponse bias analyses in its data collections, the criteria by which such methods are selected, and other available techniques for NRBA, including Bayesian methods.

Ultimately, NCES may revise its statistical standards - specifically, Standard 4.4 (http://nces.ed.gov/statprog/2002/std4_4.asp) - in light of recommendations by the Task Force.

The Task Force met in person in Washington, DC on January 17-18, 2008, and interacted by e-mail thereafter.

NATIONAL INSTITUTE OF STATISTICAL SCIENCES
TASK FORCE REPORT

NON-RESPONSE BIAS ANALYSIS

I. The Non-Response Bias Analysis Process

The focus of Task Force recommendations is unit-level rather than item-level nonresponse. The recommendations emphasize cross-sectional rather than longitudinal studies in the sense of construing nonresponse as a one-time event whose outcome is known prior to the main data collection.

The Task Force recommends that NCES follow the NRBA process shown in figure 1 and described below for all its data collections.

Steps in the Process

The principal steps in the recommended process are as follows.

1. Once nonrespondents are identified definitively, a *minimal Non-Response Bias Analysis Process (NRBA) report* is prepared, which discussed further in section 3. The minimal report compares respondents and nonrespondents on the basis of *available variables known to be predictive of responses of interest*. The minimal report can - and, the Task Force believes, should - be prepared in advance of most or all of the primary data collection.
2. If the minimal report indicates *unimportant*[2] nonresponse bias, the NRBA process proceeds to adjustment of weights for nonresponse and dissemination of the data, represented by the rightward branch in figure 1. Existing methods for weight adjustment, which typically employ some form of iterative proportional fitting to known aggregates such as state-level student populations, appear to be adequate.
3. If the minimal report indicates *important* nonresponse bias, two mandatory steps ensue, shown in the downward branch in figure 1, followed by a third stage in which up to four optional additional steps are followed. The purpose of these steps is to sharpen NCES' understanding of the nonresponse bias and to inform selection of a strategy or strategies to address it.

The *mandatory additional steps*, which can - and ideally should - occur prior to the main data collection, are:

1. Analysis of nonresponse bias at *finer granularity* than in the minimal report, but in the same manner. Specifically, this step would involve consideration of important, study-dependent subgroups. As in the minimal report, comparisons are between respondents and nonrespondents in terms of available variables known to be predictive of responses of interest.

[2] The term "important" is not precise, but is used in lieu of "significant" to emphasize that the determination may not be solely a statistical issue. In the NRBA process described here, the operationalization of unimportant is that reweighting respondents adequately alleviates the nonresponse problem.

2. *Benchmarking* the scale and nature of the nonresponse by comparison to other, related data collections. The purpose of this step is to understand whether and in what respects the nonresponse problem differs for those for the other data collections.

The *optional additional steps*, which the Task Force envisions would be selected by NCES in conjunction with the data collection contractor following the main data collection, would be chosen from:

1. *Sensitivity analysis* of major results with respect to data values associated with nonrespondents. This must be performed following the main data collection. One approach would by simulating potential values (not necessarily intended to be the "best estimates") for nonrespondents, and evaluating sensitivity of results of major interest with respect to those values.
2. *Level of effort analysis*, which can produce insight into values for nonrespondents.[3]
3. *Identification of additional predictors* of nonrespondent data values, to be used, for example, if missing responses were to be imputed.
4. *Nonrespondents follow-up*, in order to decrease the level of non-response or to obtain additional information about nonrespondents.

These four alternatives are listed in order of increasing time and financial resource requirements, and are of course not mutually exclusive.

II. Decisions by NCES

Once the mandatory and optional additional steps are performed, the Task Force envisions that NCES will make a series of decisions, which is shown graphically in figure 2. NCES would decide first, whether to discard or employ the data and second, in the latter case, whether to label nonrespondent data values as missing or to impute missing values.

These decisions are reached only when there is *important nonresponse bias*, in which case the "adjust weights" strategy recommended by the Task Force for the case of unimportant response bias is not viable. NCES would make the decisions on the basis of results of the mandatory and optional additional steps, any additional information it has available and its knowledge of the uses and users of the data.

The first decision has two principal alternatives, one of which has two principal sub-alternatives:

1. Discard[4] the data on the basis that they cannot support valid inference about the population of interest. Currently, NCES' statistical standards require this for data collections in which the response rate is too low.
2. Employ the data in NCES reports and release the data, whether publicly or via restricted data licensing. In this case, the Task Force recommends that NCES implement one of two alternatives discussed momentarily.

The Task Force acknowledges that the current policy of choosing between alternatives 1 and 2 solely on the basis of the response rate is consistent and does achieve the desired effect of preventing misuse of the data. But, the current protocol may too conservative. The Task Force believes that the NBRA process articulated here can generate sufficient information to allow NCES to employ the data even when the response rate is low, given appropriate caveats.

[3] That is, nonrespondents are likely to be similar to "hard-to-convert" respondents than to very willing ones.
[4] The term "discard" is too strong. NCES would presumably never literally discard any data. In this report, "discard" means "do use employ the data for NCES reports, and do not release the data to others."

In the event that data are to be employed, the Task Force recommends that *NCES leave (frame and any other) information about nonrespondents in the database*.[5] Doing this enables users to make their own assessments of the nature and importance of the nonresponse. The Task Force recommends that NCES then choose one of the following alternatives:

1. Employ and release the data with nonrespondent values labeled as missing.
2. Perform (multiple) imputation of nonrespondent values.

The Task Force anticipates that NCES would make the "label missing/impute" decision on the basis of the minimal report, the additional steps and other knowledge.[6] The Task Force believes that imputation is most valuable in terms of improving data quality when it can be justified scientifically and evaluated statistically. One means of evaluation is cross-validation: some respondents (in particular, those most resembling nonrespondents) are withheld from the modeling phase of the imputation, and values for them are imputed and compared to true values. When imputation is used, multiple imputation is preferable, because it permits sound assessment of imputation variance by users. The Task Force acknowledges, however, that multiple imputation is complex and that using it may prevent some users from employing the data.

The Task Force recommends the obvious: that when imputation is uses, imputed values be clearly identified, and the imputation method fully documented and released with the data. Ideally, unless it poses confidentiality issues, the code used to perform the imputation should be released, providing users the capability to replicate the process.

The Task Force believes that NCES should not make the "label missing/impute" choice on a subject-by-subject basis, although in some instances, it may be possible to make the choice on a variable-by-variable basis.[7]

Finally, the Task Force recommends that the minimal report, the results of the mandatory additional steps and the results of any optional additional steps that are performed all accompany any data release. Some editing by NCES may be necessary, of course.

III. Adjustment of Weights

The "label missing" and the "impute" alternatives require further decisions about case weights. These decisions are vital, especially because this branch of the NRBA process is associated with important nonresponse bias that cannot be mitigated solely by weight adjustment.

However, the issues are highly situation-specific, and the Task Force does not provide prescriptive recommendations.[8] Two possible default practices may be:

1) For the "label missing" alternative, if necessary, adjust weights to match to national totals, acknowledging that this is inadequate for some purposes.
2) For the "impute" alternative, retain base weights.

[5] The task force realizes that NCES may not wish to "identify" nonrepondents publicly, but it is also possible that such identification is an incentive to respond.
[6] There may also be additional considerations such as consistency with other data releases, especially those from earlier versions of the "same" study.
[7] This point is subtle because some imputed values are based on sounder evidence than others. It seems unreasonable, however, to expect data users to deal with imputed values for some but not all subjects.
[8] Perhaps the only "easy" case is when the response rate falls below NCES' standards but the bias is not important, so that standard adjustment techniques can be applied.

Difficult problems arise when multiple - and especially hierarchically structured - forms of nonresponse are present, such as nonrespondent students with a respondent school. Suppose, for example, that weight adjustment was used to handle unimportant inner-level nonresponse, but imputation was used to handle important outer-level nonresponse. Then while outer-level weights for respondents could be adjusted for inner-level nonresponse, no such adjustment is possible for outer-level nonresponse.

The Task Force recommends that weights sound enough to be used in preparation of NCES reports be included in data releases.

Figure 1: The nonresponse bias analysis process recommended by the Task Force.

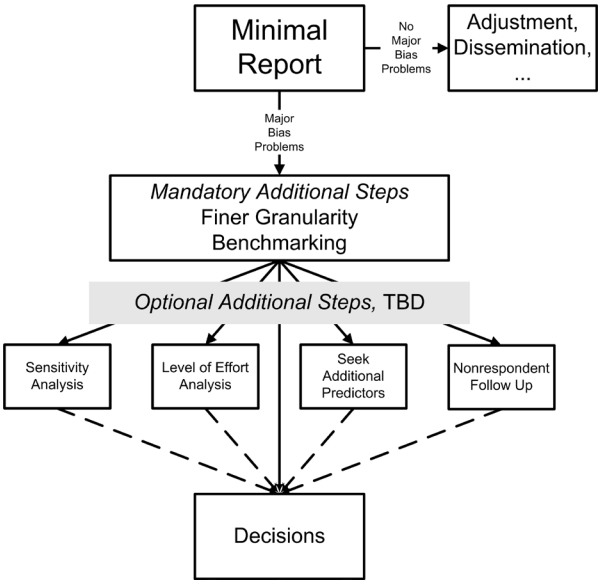
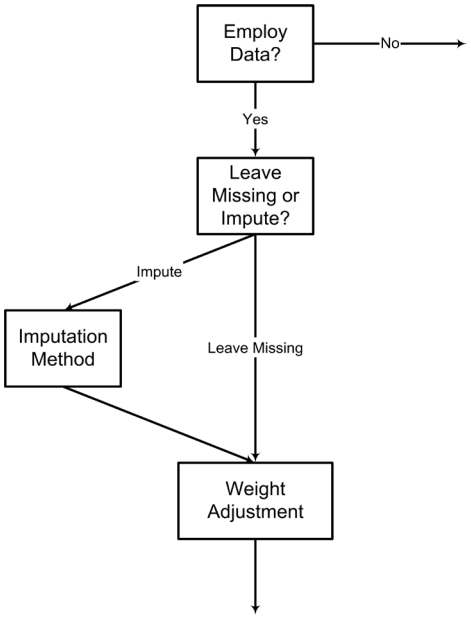


Figure 2: Decisions made by NCES following the nonresponse bias analysis process.



IV. The Minimal Non-Response Bias Analysis Report

Since the minimal report introduced in section 1 is a new concept, we both describe it in some detail and present an illustrative example in appendices A and B.

In the minimal report, respondents and nonrespondents are compared with respect to *available variables known to be related to important responses*. This is a central point: the rationale is that observed differences between respondents and nonrespondents are important principally in the extent to which they affect important responses in the data, such as student performance.

The Task Force emphasizes the difference between *available variables* and *frame variables*. Many of the presentations to the Task Force at its January 2008 seemed to indicate that respondents and nonrespondents were compared only in terms of frame variables even when the set of available variables was much larger.[9] The Task Force finds that this practice is neither necessary nor justified. Additional variables may be available by linking to other NCES or external databases, although the latter may introduce data quality problems.

To illustrate, assume that the data collection is a cross-sectional survey that is repeated over time, so that there is a "most recent predecessor" of the one under consideration, and that the sampling frame is an NCES universe data collection - CCD, PSS or IPEDS.[10]

The minimal report is based on the data schema shown in figure 3. Attributes fall into three classes, with two superclasses.

- **Universe Variables:** Those in the sample frame-associated universe data collection, which we label X_{U}. Within X_{U} we distinguish.
  - **Sampling Variables:** The subset of variables in X_{U} used to draw the sample and to form the base weights, denoted by X_{samp}.
- **Survey Variables:** Those collected by the survey, only from respondents, which we denote by Y and partition into
  - **Responses:** Variables in Y known by means of domain knowledge to be responses of interest for scientific or policy purposes, denoted by Y_{resp}. **Predictors:** Variables in Y ordinarily used as predictors for the $Y_{\mathrm{resp,}}$ which we denote by Y_{pred}.[11]

Data availability is then[12] *X* for both respondents and nonrespondents, but Y only for respondents.

Universe variables *X*		Survey variables *Y*	
Sampling variables X_{samp}		Predictors Y_{pred}	Responses Y_{resp}

Figure 3: The data schema assumed in the discussion of the minimal report.

[9] For instance, the available variables may be all those in the CCD or PSS.

[10] Of the programs about which presentations were made to the task force on January 17, 2008, only NHES does not fit this model at the institution level.

[11] More simplistically, Y_{pred} consists of all variables in *Y* other than those in Y_{resp}.

[12] This is an "all-or-nothing" model that is not valid in all cases, and especially not in longitudinal data collections.

The Task Force recommends that the following steps be performed in order to prepare the minimal report, each of which has associated issues.

1. First, using X (not merely X_{samp}!) and Y from *respondents in the most recent predecessor*, determine which variables in X are statistically effective - and, if possible, scientifically meaningful - predictors of the variables in Y_{resp}, which are denoted generically by X_{pred}. Note that some of these may not be in $X_{\text{samp.}}$
2. **Issues:** This is a model selection problem, so how is the selection done? Are the Y_{resp} to be predicted jointly or individually? How sensitive is the result to the choice of the Y_{resp} and the model selection procedure? How should Y_{pred} be treated in this process?
3. Compare respondents and nonrespondents on the current survey in terms of the values of the X_{pred} at a relatively low level of granularity (high level of aggregation), report the differences, and determine if they are important.
4. **Issues:** What level of granularity? How is the comparison performed? What defines "important," and how sensitive is the process to its specification?

Concerning techniques used to compare respondents and nonrespondents, the Task Force observes that, statements in NCES Standard 4-4 notwithstanding,[13] the prevailing practice (see appendix B) is to compare variables one-at-a-time, using t-tests for numerical variables and chi-squared tests for categorical variables. For comparability, the same was done in the experiment described in appendix A. However, we show in appendix B that it is straightforward to use multivariate machine learning techniques (specifically, recursive partitioning, a form of classification) to compare respondents and nonrespondents simultaneously on multiple (mixed categorical and numerical) variables.

The Task Force recommends that for minimal reports, NCES encourage more forcefully the use of multivariate methods, which would have the further benefit of reducing multiplicity issues. Given the ready availability of software implementing these techniques (see appendix B), there is little case for not using them. Depending on the extent to which it wishes to be prescriptive about specific methods, NCES may wish to explore the potential of several supervised machine learning methods (i.e., classification) such as random forests and support vector machines, which can cope more readily than formal, analytical multivariate models with high-dimensional data, and often involve fewer distributional assumptions. The CHAID technique used to assess nonresponse bias in NCES (see appendix C) studies is a step in this direction.

In a private communication entitled "On the role of proxies for survey outcomes in the analysis of nonresponse bias," Task Force member Roderick Little lays out a related path using proxies for the "available variables known to be related to important responses."

V. Other Issues

The Task Force emphasizes that the approach laid out in sections 1-3 seems better suited to repeated cross-sectional studies than to longitudinal studies, because it conceptualizes nonresponse as an "all-or-nothing" phenomenon. Attrition in longitudinal studies is a ubiquitous, difficult phenomenon that does not fit neatly within the framework in section 1. An ongoing series of minimal reports could be used to assess the effects of attrition. A default strategy of "retain pre-atttrition data and label missing post-attrition data" may be a defensible approach.

[13] For instance, Guideline 4-4-2B states "*Formal multivariate modeling* can be used to compare the proportional distribution of characteristics of respondents and nonrespondents to determine if nonresponse bias exists…" (emphasis added).

Nor does this report account fully for interactions between the NBRA process described in section 1 and actions undertaken by contractors on behalf of NCES to reduce nonresponse.

Finally, to acknowledge the obvious, the high-level description of the Task Force-recommended NRBA process does not capture many of real-world, situation-specific complexities faced by NCES. The Task Force recognizes this, but at the same time, believes that its recommendations constitute a sound set of principles on which NCES can base a realistically sensible implementation.

APPENDICES

A. Experimentally Constructed Minimal Report

B. Using Machine Learning in the Minimal Report

C. Summary of NRBA Procedures for Surveys Presented

D. Expert Panel Members

Appendix A: Experimentally Constructed Minimal Report

Important questions about the minimal report are "Is the process feasible?" and "Are the results informative and actionable?" NISS has conducted experiments showing that it is feasible, and not onerous. Actionability is, ultimately, a decision that can only be made by NCES.

The experiments compare key variables between respondents and nonrespondents, with appropriate weights applied. Often the first step is to select key variables from a large set of potential predictors of the responses of interest. When manageable, it may be desired to compare all available variables. Potential explanatory variables not available for respondents and nonrespondents should also be identified.

Two approaches have been studied. Steps in **Approach 1** are:

1. Identify a subset of variables available for both respondents and nonrespondents that are significantly associated with key responses.

   Using data for respondents in the most recent predecessor dataset and frame data $\mathrm{X_U}$, a subset of $\mathrm{X_U}$ is identified as "significant predictors" of key response variables $\mathrm{Y_{resp}}$, using a statistically justifiable variable selection approach. The Task Force does not make a specific recommendation of how predictors should be selected. Note, however, that the goal is not to find a parsimonious model and scientific judgment may play a role. An underlying assumption is that the significant predictors of the responses are the same for respondents as for nonrespondents.

2. Compare respondents and nonrespondents on the variables selected in step 1. Typically (see appendix C), *t*-tests are used for continuous variables and chi-square tests for independence used for categorical variables, with a standard significance level of .05. At this stage base weights are used for analyses.

3. Note any variables found to be significantly different between respondents and nonrespondents.

Steps in **Approach 2** are:

1. Compare respondents and nonrespondents on all variables available for both groups, from frame and/or previous study. Typically, *t*-tests are used for continuous variables and chi-square tests for independence used for categorical variables, with a standard significance level of .05. At this stage, base weights are used for analyses.

2. For variables found to be significant, assess whether they are significant predictors of the key outcome variables of the survey. Further evaluation and adjustments will be necessary.

In general, Approach 1 is preferred, because the results of the variable selection process can be assessed for scientific plausibility as well as statistical efficacy.

The following experimental example is based on the 1999-2000 first grade students and school's data from ECLS-K. The data for respondents and nonrespondents is taken from the CCD and PSS.

For this example, approach 2 was used. A "most recent predecessor" data set is not available so the data itself were used in this role

The key responses are spring mathematics and reading scores. The candidate predictors from the frame include public/private, school type, full-time-equivalent teachers, lowest grade in school, highest grade in school, number of students in first grade, total students in school, percent of students in school who are Asian American, American Indian, Hispanic, and Black, student-teacher ratio, percent of students in

school who are male, and Census urbanicity category. Variables available for public schools only include indicators for Title 1 eligibility, magnet school, or charter school, number of students eligible for free and reduced-price lunches, and number of migrant students. For private schools, religious affiliation (Catholic, other, none) was also used. Base weights were not available for nonrespondents, so for illustrative purposes, first grade enrollment is used as a proxy for weight.

1. Compare respondents and nonrespondents on all variables available for respondents and nonrespondents.

   The variables with significant differences, based on tests of association for categorical variables and t-tests for continuous variables, are lowest and highest grade offered, grade 1 enrollment, and student teacher ratio. Of these the last two seem more likely to affect student performance. The results are summarized below.

Table 1: Categorical Variables

Variable Name	Chi-square	p
Type	67	<.0001
Public	82	<.0001
Title I (public)	63	<.0001
Magnet (public)	135	<.0001
Charter (public)	259	<.0001
Religious Aff. (private)	1719	<.0001

Table 2: Numerical Variables

Variable Name	T	p
FTE teachers	1.05	.2952
Lowest grade	0.92	.3570
Highest grade	-0.75	.4545
Grade 1 Enrollment	3.14	.0017
Total Enrollment	0.13	.8981
American Indian %	-0.32	.7511
Asian American %	1.55	.1210
Hispanic %	0.31	.7579
Black %	-0.22	.8254
Student Teacher Ratio	-0.80	.4211
Male %	-0.86	.3923

2. The variables found to be significantly different between respondents and nonrespondents are School Type, Public, Title I, Magnet, Charter, Religious Affiliation, and Grade 1 Enrollment. Whether or not these are significant predictors of the outcome variables should be included as part of the assessment of nonresponse bias. All of these variables are found to be significant predictors of Reading and/or Math scores.

Appendix B: Using Machine Learning in the Minimal Report

As noted in section 3, Guideline 4-4-2B of the NCES Statistical Standards states that "Formal multivariate modeling can be used to compare the proportional distribution of characteristics of respondents and nonrespondents to determine if nonresponse bias exists…". The Task Force understands that "can" is different from "should," but believes that today multivariate comparison is neither difficult to perform nor difficult to interpret, and should be done if possible.[14]

The factors underlying this view have changed since the Statistical Standards were last revised in 2002. Formal (in the sense of analytical - fitting presumably nonparametric multivariate distributions and performing explicit tests of hypotheses) comparison remains problematic in terms of implementation and interpretation. However, the rapid development and widespread availability of methods for supervised[15] machine learning (also known as *classification*) make such comparisons rather straightforward.

Moreover, modern classification techniques can readily handle:

- Very large data sets, although this is not an issue for NCES surveys;
- High-dimensional data (dimensionality is a clear impediment to use of classical methods);
- Data containing both categorical and numerical data.

Therefore, the Task Force urges that NCES be more insistent regarding use of multivariate techniques to compare respondents and nonrespondents.

To illustrate, NISS repeated the experiment described in appendix A using the same data and the (recursive) partitioning capability of the software package JMP®.[16] The result is a tree, as shown in figure 4, that shows how the data elements "split" with respect to respondent (blue) vs. nonrespondent (red):

- The first, and "most informative" split is with respect to the variable RELIG2, whose values are 0 = public, 1 = catholic, 2 = other, 3 = (private) nonsectarian. The *nonresponse* rate is significantly higher among schools for which RELIG2 is 2 or 3 than among those for which it is 0 or 1.
- Within the subset "RELIG2 = 0 or 1," the most informative split is between the two values 0 and 1, with nonresponse higher when RELIG2 = 0. However, within the subset "RELIG2 = 2 or 3," the most informative split is on the value of GSHI (the highest-grade present in the school), and specifically whether GHSI is less than 10 (lower nonresponse rate) or at least 10 (higher non-response rate).
- With the subset "RELIG2 = 2 or 3, GSHI < 10," the most informative split is on PUPTCH (pupil/teacher ratio): nonresponse is high when PUPTCH < 6.909.

No other splits were determined to be informative.[17]

[14] Rather than, as "can" suggests, if desired.

[15] In broad terms, "supervised" machine learning applies to data containing a "response," which for NCES surveys is respondent as opposed to nonrespondent. "Unsupervised" machine learning pertains to grouping (clustering) data on the basis of similarity of characteristics. Both are often used for predictive purposes, following a "training" phase using initial data. This aspect is not immediately relevant, but it could be used to inform the design and execution of future surveys on the basis of a predicted likelihood of response.

[16] A product of the SAS Institute.

[17] The analysis was more subtle than this description indicates in the sense that variable transformations were needed in order to eliminate spurious effects. In particular, all sub-school pupil counts were replaced by ratios with respect to total enrollment.

Note that this analysis reveals a three-way interaction (among RELIG2, GSHI and PUPTCH) that would not have been identified had only single-variable comparisons been performed and would have been hard to detect by analytical modeling.

The total time required to do the entire analysis was less than one day, most of which was devoted to transforming variables. Figure 4 is a screen capture of the output produced by JMP®.

Figure 4: Tree produced by JMP for ECLS-K nonresponse bias analysis.

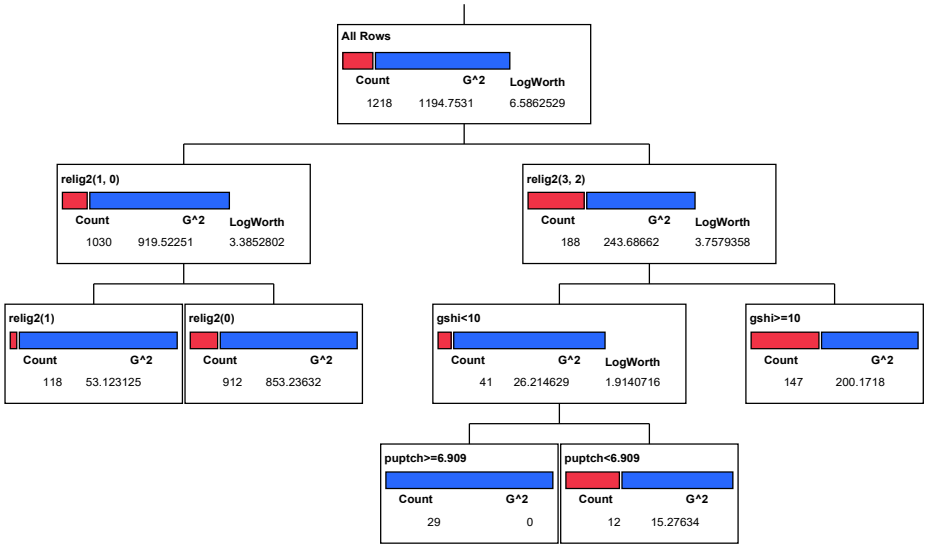
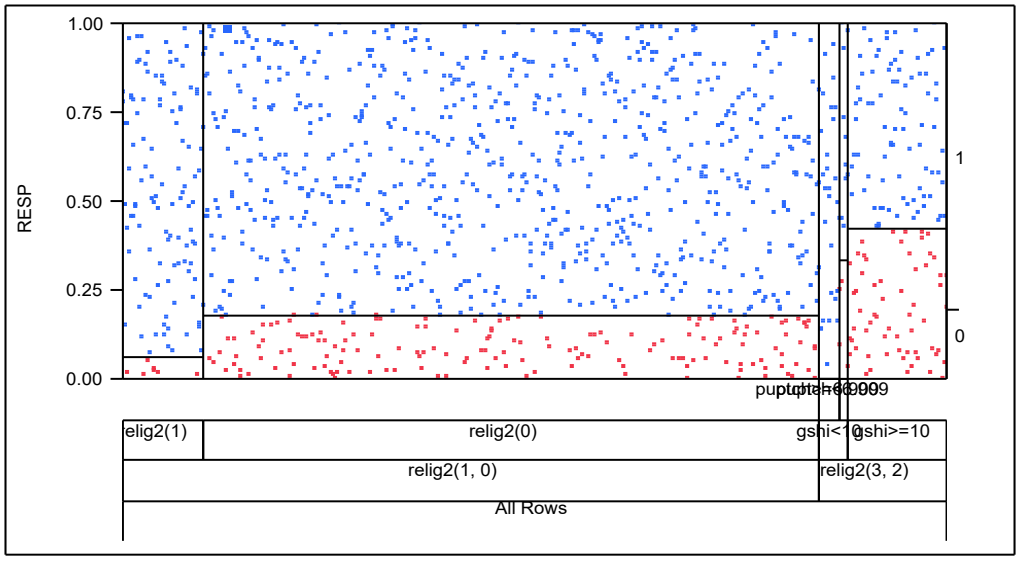


Figure 5: Alternative representation of results of partitioning ECLS-K data.



Appendix C: Summary of NRBA Procedures for Surveys Presented

	NHES	NAEP	PISA/PIRLS	SASS	NPSAS	ECLS-K	ELS
Frame	Eligible House-Holds		CCD+PSS	CCD+PSS	IPEDS	CCD+PSS	CCD+PSS
X's used to generate sample	RDD		MSA, county, # eligible students	BIA, State, grade level, # teachers, pvt. sch cat.	Control, level, Carnegie, state, size. Student type.	Census region, MSA, race/eth, MOS, pc income	Census div/regn, urbanity, # 10th graders. Ethnicity
How X's chosen	CHAID	Historic information	Logistic model w/stepwise	Judgment	CHAID	CHAID	
Basis for bias measure	Level of effort	Full sample	Full sample	Frame		Level of Effort	Full Sample
Comparison Method		Chi/t	Chi/t	Chi/t			Chi/t
External Data	CPS, past, Census	transcript			Pell Student Loans	NHES	
Substitutes/ Replacements		Substitutes	Substitutes				
Multiplicity Adjust					Yes		
Sensitivity Analysis	For items with <90% response						
Additional Steps	2007 Bias Study						Survey NR schools

C.1 NHES – National Household Education Survey

Overview: Covers learning at all ages, from early childhood to school age through adulthood. Data collection is via household/landline telephones. Advance mailings, follow-up conversion letters, and some incentives used to increase response rates. Since 2003, sub-sampling of screener nonrespondents for extensive follow-up.

Assessment: Profile response rates by subgroup using frame data (Census) and check consistency of patterns with previous studies (e.g. Groves & Wissoker). Sensitivity analysis for items w/nonresponse < 90%.

External Data: Compare to CPS, previous NHES studies on demographic (age, age by grade, race/ethnicity, family structure, employment status) and key survey variables (participation in care arrangement, parental involvement, adult education participation).

Bias Measure: Compare reduced effort and full effort (weighted) key survey estimates within subgroups. Few statistically or practically significant differences noted. (presumably t-test, chi-square).

Adjustments: Use CHAID analysis to identify cells (w/in phone exchange) for weighting class adjustment for nonresponse. Continuous variables are categorized by decile.

Explanatory Variables: Use those delivered by MSG. Relate x's to response propensities. Variables available in frame for both screener respondents and nonrespondents are only at telephone exchange level. Predictors used include:

p(respond) for screener: % white, Census division/region, % hispanic, median home value, % with income > 70K, % homeowners, MSA.

(respond) for students: age, grade home school, Census region, urbanicity (ECPP, ASPA); if student responded to screener, sex, education level, education participation (AE).

Item Nonresponse: Used imputations and compared estimates using observed and completed data. Conducted an extreme-imputation analysis establishing potential for NRB though did not find actual evidence of bias.

Reporting/Results: Reported on response rates and gave CHAID results in tabular form**.** Additional considerations for interpretation of results mentioned. Concluded that there was no substantive bias and additional efforts to increase response rates did not affect nonresponse bias.

Comments

- NHES 2007 Bias Study in progress.
- Potentially more explanatory variables not being used.
- Have not looked at measures of fit.
- NHES 2001 report indicates use of hot-deck imputation.
- Sensitivity analysis conducted for items with <90% response.

C.2 NAEP - National Assessment of Educational Progress

Overview: Biennial survey with separate samples for grades 4, 8, and 12. Grade 12 results reported separately by public and private. Private schools are main source of NRB. Substitute schools may be used. Student response rate is high except for excluded students and seniors. Methods described for 2007 survey.

Assessment: Response rates are computed for each combination of grade and public/private. NRBA conducted for 4 groups with < 85% response rates.

External Data: Occasional transcript studies.

Bias Measure: Compare distribution of weighted original full sample to base-weighted responding schools, w/ and w/o substitutes, and to final sample with adjusted weights. Use t-tests and R-S chi-square tests. NRB for schools based on ethnicity distributions and enrollment; for based on gender, race/ethnicity, age, NSLP eligibility, SD, ELL.

Adjustments: Weights are adjusted. Selection of characteristics used not described.

Explanatory Variables: Have a good deal of frame data about schools and historical data on which variables are related to outcome variables. Limited data on students. Variables used include:

p(respond) for schools: Census region, private school subgroup, type of location, enrollment, ethnicity distribution.

p(respond) for students: Sex, race/ethnicity, relative age, NSLP eligibility, SD, ELL.

Item Nonresponse:

Reporting/Results: Tabulate results by subgroup. Include bias, relative bias, and p-values from t-tests and chi-square tests. Found adjustments and substitutions at least partially effective.

Comments

- Past NRBA have used logistic regression to consider joint effect of all variables on response propensity.

C.3 PISA/PIRLS – Program for International Student Assessment/Progress in International Reading Literacy Study

Overview: PISA measures cumulative knowledge of 15-year-olds and PIRLS measures reading comprehension of 4th graders. Both are USA portions of international surveys. Two-stage stratified PPS samples taken from CCD (Common Core of Data) + PSS (Private School Universe Survey). Substitute schools used (note pres. Says replacement but discussion indicated substitutes). Monetary incentives for students are successful.

Assessment: Overall response rates computed for each sample.

External Data: None indicated.

Bias Measure: Compare distributions (at school level) of respondent schools and original sample for variables school control, community type, Census region, poverty level, enrollment, race/ethnicity distribution, % NSLP eligible. T-tests and R-S chi-square tests used. Student nonresponse is less of an issue, though specifics of response rates are not given and no bias assessment was done.

Adjustments: Weighted by design weights (school size) and eligible students. Calculations in SUDAAN.

Explanatory Variables: Not sure how they are chosen. For schools, it appears all frame variables available for all schools are used.

Item Nonresponse:

Reporting/Results: Tabular results by public/private indicate some NRB with respect to race and public/private. Over/under-representation noted for a few subgroups. Reported bias, relative bias, and p-values for mean enrollment for PIRLS and response rate and standard error for PISA.

Comments

- Little data on nonresponding students; analysis was only done at the school level.
- NRBA conducted for 120 responding schools then repeated for 120 + 63 replacements/ substitutes.
- Age-based complicates efforts, less clout than NAEP.
- Stepwise selection with logistic model predicting response propensity described in report.

C.4 SASS – Schools and Staffing Survey

Overview: Survey of schools, districts, principals, teachers, and library media centers. Uses stratified PPS sampling, where strata used change over time, using CCD + PSS.

Assessment: Compute response rate for each of twelve data files, determined by sector and respondent type. Similar for item bias analysis. For data files with unit response < 85%, identify subpopulations with base-weighted response rate < 85%.

External Data: None described.

Bias Measure: Compare base-weighted respondents and final-weighted respondents to frame within selected reporting characteristics to identify bias for items not well-correlated with weighting factors and determine if bias has been masked for items correlated with weighting factors. Noteworthy differences defined as > 10% difference relative to frame proportion, 1% absolute difference, $r < .15$, and cell has at least 30 interviews.

Adjustments: Not described.

Explanatory Variables: Selection not described. (may have said judgment)

Item Nonresponse: Similar to unit NRBA. Hot-deck and other imputation used.

Results/Reporting: Adjusted distributions (proportions), standard errors, and t-statistics reported in tables.

Comments

- Some problems w/standard errors and proportions in tables, where standard errors reported for "Frame distribution (adjusted for ineligible units and standard error)".

C.5 NPSAS – National Postsecondary Student Aid Study

Overview: Quadrennial survey of students at schools eligible for Title 4. Sample frame is IPEDS. Initial survey is conducted online, with contact via mail and email. Follow-up surveys may be conducted over the phone. Monetary incentives provided during the initial period. Additional school's data available in IPEDS and additional student data is obtained from schools and other databases such as student aid applications. Schools may be reimbursed for record abstraction.

Assessment:

External Data: None mentioned.

Bias Measure: Bias compared before/after imputations for known variables: region, institution total enrollment, CPS match, Pell Grant recipient, Stafford Loan recipient, and Pell/Stafford amounts.

Adjustments: Item nonresponse bias – weighted sequential hot deck imputation used where interviews not completed. CHAID analysis was used to identify imputation classes, plus practical considerations determined by subcontractor. Weight adjustments computed using propensity models in 3 stages: unable to locate, refusal, other nonresponse types. ROC curve to assess (Wilcoxon test used to test predictive fit).

Explanatory variables: Used CHAID to determine interaction terms. Main effects include:

p(respond) for schools: School type, Carnegie classification, OBE region, HBCU status, % aid (by aid type), %ethnicity (by group), % male, and % graduate/first professional enrollment.

p(respond) for students: region, institution enrollment, CPS match, Pell and Stafford loan amounts. % full-time Fall enrollment and in-stage tuition also used; however, these are not available for nonrespondents so school-level IPEDS data used.

Item Nonresponse: Used predictors from student NRBA as well as gender and group which were used because they were known, and institution strata. Use carefully constructed imputation classes, donor-imputee matching criteria, and random hot-deck searches within imputation cells to reduce NRB. Bias equal mean before imputation minus mean after imputation.

Results/Reporting: Tabular reports give mean and median "estimated bias" for different groups as well as % significant bias.

Comments

- Multiple imputation recommended by audience for efficiency gains.
- Only 2-year public institutions had student response rate below 85%, so student NRBA only conducted for this group. Adjusted for different types of nonresponse.
- Use ROC curve to assess overall predictive ability of nonresponse model.

C.6 ECLS-K Early Childhood Longitudinal Study – Kindergarten Cohort

Overview: Data collected in K, 1, 3, 5, 8. After base year, sample loss occurs due to children changing residences and/or schools. Some attempt to follow movers and subsampling used in some waves. The sample is refreshed in 1st grade only.

Respondents can drop out one year and return to sample.

Assessment: Compare rates to population estimates from frame, other surveys (NHES). Simulated nonresponse based on patterns from external instruments. Compared base and various adjusted weights, current year weights.

External Data: Survey estimates compared to NHES.

Bias Measures: Evaluation of sample attrition – weighted estimates of bias – base year, wave 6; also, mover subsampling; take into account variance of the bias. Evaluate effect of longitudinal attrition, bias due to mover subsampling, effectiveness of sample-based raking in reducing variability due to small sample sizes due to design and sample loss.

Adjustments: For base year, inverse probability weights adjusted for nonresponse. Wave 6 weights adjusted for subsampling of movers, nonresponse, raking. Up to 40 longitudinal weights adjusted.

Explanatory Variables: Use CHAID to examine relationships between respondents and nonrespondents on variables known from frame. Variables used for CHAID analysis include Census region, school affiliation, type of locale, total enrollment, and % non-white.

Item nonresponse: Profiled response rates by type.

Results/Reporting: Reported estimates of weights, relative bias, and RMSE. Plotted ratios of RMSE raked and unraked estimates using cross-sectional and longitudinal weights to evaluate efficacy of removing bias (ratio >1). Found sample-based raking to be effective in reducing bias and associated variance estimates. Report includes CHAID diagrams illustrating relationship between school characteristics and response rates.

Comments

- Difference between movers moving for different reasons not accounted for.
- Base-year weights used to assess bias; potentially underestimated.

C.7 ELS - Education Longitudinal Study of 2002

Overview: Survey of high school sophomores with sample freshening in grade 12. Sample based on students but includes interviews with students, teachers, parents, and administrators. Only base year nonresponse was discussed.

Assessment: Overall response rate computed for schools and students. School response bigger problem.

External Data: Supplemental nonrespondent survey.

Bias Measures: Compare to full sample estimates. Estimate bias using design-based weights. Re-evaluate using nonresponse-adjusted weights. All variables available for respondents and nonrespondents used. Variance of bias estimated using SUDAAN.

Adjustments: Use known and high-response variables as proxies for unknown. Models predicting response propensities used.

Explanatory Variables: Selection process not clear but it appears some selection process was used, and also CHAID. For nonresponding schools, data on school characteristics collected. For nonresponding students, only sex and race/ethnicity were available, along with school-level variables.

Item nonresponse: Imputed missing values, multiply in some cases using PROC MI. Potential magnitude of bias measured as nonresponse rate times difference in characteristic values between respondents and nonrespondents. Known characteristics include school type, MSA or urbanicity, and Census region. High-response characteristics used for students include sex and race/ethnicity. Other variables hypothesized to be helpful in explaining nonresponse are mother's education, language minority status, reading quartile, and math quartile.

Results/Reporting: Plotted estimated relative bias before nonresponse adjustment vs. after adjustment, for school-level and student-level bias. Found several small but statistically significant biases using t-tests and chi-square tests on several variables that were eliminated with nonresponse adjustments. Additional results in report include detailed tabular summaries of estimates, bias, and relative bias, as well as a plot of Type I error rate vs. bias ratio. Also, extensive summaries of item response bias.

Comments

- Suggestions included to use predicted outcome as proxy and to examine correlation between proxies and targets. Also comments on results relative to propensity score adjustments.
- They said they did not assess reducing bias with imputation but report does mention use of weighted hot-deck imputation, as well as PROC MI, for the purposes of reducing bias.

Appendix D: Expert Panel Members

David Cantor, WESTAT

Martin Frankel, CUNY

Robert Groves, University of Michigan

Brian Harris-Kojetin, OMB

Nancy Kirkendall, Consultant

Frauke Kreuter, University of Maryland

Roderick Little, University of Michigan

Panel convened by National Institute of Statistical Sciences

Alan Karr, NISS

Satkartar (Saki) Kinney, NISS