NSAF

2002 NSAF Data Editing and Imputation

Report No. 10

Prepared by:

Timothy Triplett

Assessing
the New
Federalism
An Urban Institute
Program to Assess
Changing Social Policies

Methodology Reports

PREFACE

2002 NSAF Data Editing and Imputation is the tenth report in a series describing the methodology of the National Survey of America's Families (NSAF). One component of the *Assessing the New Federalism* project at the Urban Institute and conducted in partnership with Child Trends, the NSAF is a major household survey focusing on the economic, health, and social characteristics of children, adults under the age of 65, and their families. Westat conducted data collection for the survey.

During the third round of the survey in 2002, interviews were conducted with over 40,000 families, yielding information on more than 100,000 people. The survey sample is representative of the nation as a whole and of 13 focal states, and therefore allows for both national as well as state-level analysis.

About the Methodology Series

This series of reports has been developed to provide readers with a detailed description of the methods employed to conduct the 2002 NSAF. The 2002 series includes the following reports:

No. 1: An overview of the NSAF sample design, data collection techniques, and estimation methods

No. 2: A detailed description of the NSAF sample design for both telephone and in-person interviews

No. 3: Methods employed to produce estimation weights and the procedures used to make state and national estimates for *Snapshots of America's Families*

No. 4: Methods used to compute and results of computing sampling errors

No. 5: Processes used to complete the in-person component of the NSAF

No. 6: Collection of NSAF papers

No. 7: Studies conducted to understand the reasons for nonresponse and the impact of missing data

No. 8: Response rates obtained (taking the estimation weights into account) and methods used to compute these rates

No. 9: Methods employed to complete the telephone component of the NSAF

No. 10: Data editing procedures and imputation techniques for missing variables

No. 11: User's guide for public use microdata

No. 12: 2002 NSAF questionnaire

About This Report

Report No. 10 focuses on the data editing techniques and imputations that were unique to the 2002 NSAF data processing steps. It is a supplement to the 1997 and 1999 NSAF data editing reports (No. 10 in both series), and does not reiterate the data editing techniques, data processing, and coding guidelines documented in these prior reports.

For More Information

For more information about the National Survey of America's Families, contact:

Assessing the New Federalism
Urban Institute
2100 M Street, NW
Washington, DC 20037
E-mail: nsaf@ui.urban.org
Web site: http://anf.urban.org/nsaf

Tim Triplett

CONTENTS

OVERVIEW OF DATA EDITING AND ITEM IMPUTATION

Introduction

The National Survey of America's Families (NSAF) was developed out of a need for empirical data to respond to major changes in welfare policy at the federal and state levels. The Urban Institute and Child Trends combined their expertise in welfare policy with the data collection efforts of Westat, an experienced survey firm.

By the third (2002) round of NSAF, many of the data handling problems encountered in the first two rounds of data collection had been minimized, and proven data handling approaches were implemented with few changes. This report describes those procedures unique to the third round. To read about the data editing efforts of rounds one and two please see Report 10 in the 1997 and 1999 NSAF Methodology Series.

Data Editing and Data Coding

The data editing process for the 2002 NSAF consisted of three main tasks: handling problem cases, reading and using interviewer comments to make data updates, and coding questions with text strings. Extensive quality control procedures were implemented to ensure accurate data editing and coding.

Before delivering the data, Westat did much of the data editing involving interviewer comments as well as fixing cases that were designated as problematic by the telephone supervisor. Additionally, the Urban Institute staff developed a series of programs that checked the consistency of each question item by section of the questionnaire as well as general program checks that tested for item consistency on subjects such as family relationships, immigration status, and age of respondents. These section and general program checks often uncovered problems that required going back to the interviewer comments or problem case notes in order to resolve.

Except for the coding done at the Census Bureau of industry and occupation, all the post-survey editing and assignment of codes for open-ended questions for the 2002 NSAF was done at the Urban Institute. "Other specify" questions were those in which a question had some specific answer categories but also allowed text to be typed into an "other" category. Open-ended questions had no pre-coded answer categories. Westat and the Urban Institute had developed an interactive process for defining these categories during round 1. It was this structure that formed the basis for much of the coding done for the 2002 survey. Often, for "other specify," we were able to start with the exact decisions made in round 1 for a respondent comment.

Because data editing and coding update the data, careful quality control procedures were implemented at Westat and at the Urban Institute. These measures involved limiting the number of staff who made updates, using flowcharts to diagram complex questionnaire sections, frequent consulting meetings, carefully checking updates, and conducting computer checks for inconsistencies or illogical patterns in the data.

Data edits and open-ended and other specify coding make up the bulk of the data development work needed to produce the analytical files. However, this oversimplifies the work needed to produce analytical files. At the end of this report is a flow chart that displays how highly structured the overall NSAF data development process was. This diagram starts with the raw interview data and ends with nine analytical public use data files. In addition to coding and editing the NSAF data development process relies on imputation methods to account for much of the item missing data.

Imputation

For most NSAF questions, item nonresponse rates were very low (often less than 1 percent), and seldom did we impute for missing responses. The pattern and amount of missing data from round to round varied very little, enabling us to use similar imputation approaches across all three rounds. The answers to opinion questions were not imputed for any cases where they were missing. Still, there were important questions for which missing NSAF responses were imputed to provide a complete set of data for certain analyses. For example, the determination of poverty status is crucial, but often at least one of the income items that had to be obtained to make this determination was not answered. For every variable where imputations were made there is a corresponding variable on the same dataset that indicates the imputed observations.

As is the case in many household surveys, the NSAF encountered significant levels of item nonresponse for questions regarding sensitive information such as income, mortgage amounts, health care decisions, and so forth. In fact, the income-item nonresponse could range up to 20 or even 30 percent in the NSAF. Hence, the problem could not be ignored. The imputation of missing responses is intended to meet two goals. First, it makes the data easier to use. For example, the imputation of missing income responses permits the calculation of total family income (and poverty) measures for all sample families—a requirement to facilitate the derivation of estimates at the state and national levels. Second, imputation helps adjust for biases that may result from differences between persons who responded and those who did not.

The "hot deck" imputation (e.g., Ford 1983) approach was used to make the imputations for missing responses in the NSAF. In a hot deck imputation, the value reported by a respondent for a particular question is given or donated to a "similar" person who failed to respond to that question. The hot deck approach is the most common method used to assign values for missing responses in large-scale household surveys.

The first step in this hot deck imputation process was separating the sample into two groups: those who provided a valid response to an item and those who did not. Next, a number of matching "keys" were derived, based on information available for both respondents and nonrespondents. These matching keys vary according to the amount and detail of information used. One matching key represents the "highest" or most desirable match and is typically made up of the most detailed information. Another matching key is defined as the "lowest" or least desirable. Additional matching keys are defined to fall somewhere between these two; when combined, these keys make up the matching hierarchy.

The matching of respondents and nonrespondents for each item is undertaken based on each matching key. This process begins at the highest (most detailed) level and proceeds downward until all nonrespondents have been matched to a respondent. The response provided by the donor matching at the best (highest) level is assigned or donated to the nonrespondent. For the most part, respondents are chosen from the "pool of donors" without replacement. However, under some circumstances, the same respondent may be used to donate responses to more than one nonrespondent. By design, multiple uses of donors are kept to a minimum. An imputation "flag" is provided for each variable handled by the imputation system. In fact, all imputations assigned can be easily tied to the donor through the Donor ID number, because it is retained. The linkages between donor and donee were all kept as part of the complete audit trail maintained throughout the process, although they are not currently being made available on the NSAF Public Use Files.

Since the hot deck procedures used for imputing missing data for the 2002 NSAF were the same procedures used in the previous two rounds of data collection we will not detail them here. Please refer to either the 1997 or 1999 data editing and imputation NSAF methodology report No 10 for more information about the hot deck imputation methods used.

Imputation procedures for the telephone questions (M14 through M22)

The survey questions to determine the number of residential telephone numbers in a household were changed in the 2002 questionnaire. This new data collection sequence also required new imputation procedures for missing phone line data that the Urban Institute imputed internally. Determining the number of residential telephone numbers within a household is important for determining a household's probability of selection. The two previous rounds of the NSAF (1999 and 1997) used a simple two-question approach to estimate total residential phone numbers. The 2002 questionnaire asked each household a more complex set of questions that not only collected the number of residential telephone lines but also sought to determine how each telephone line was used (residential usage, business usage, fax/Internet line only). Overall, most respondents answered the telephone questions, but more than 400 households did not. Most of the missing data resulted from respondent break-offs, in which the respondents completed enough of the NSAF interview (at least through section K) to be included as part of the final data set but did not get asked the telephone questions at section M. Since the characteristics of persons who break off early in a survey differ from those who do not, a decision was made to impute the missing data using household characteristics that are deterministic of multiple telephone lines.

Figure 1 shows how the telephone questions were asked on the 1997 and 1999 rounds of the NSAF compared with the 2002 version. There were more questions in 2002 and the wording of questions M14 and M15 was changed. In 1997 and 1999 interviewers were instructed not to include cell phone numbers in M14 and were instructed to include home computer fax numbers in M15 if they were also used for voice communication. Therefore, while the questions used were different, the goal of estimating total residential phone numbers within a household that could be used for voice communication was the same.

In total, 466 households required imputations for one or more of the telephone questions. Of the 466 households, 403 were a result of completed interviews that broke off before question M14. In addition, 24 respondents refused to answer M14, thus, 427 of the 466 households that received imputed values needed to have the first question in the sequence (M14) imputed.

There were also 13 households that answered “don’t know” to the first question about additional phone numbers (M14). The 2002 study asked these 13 households a follow-up question to determine whether there were additional numbers for computer of fax machines. If there were additional numbers for fax or computer lines, the respondent was asked how many (M19), subsequently returning to the normal questionnaire sequence. So, for the respondents who did not know the answer to question M14, there was enough information obtained to deterministically edit the value of M14 and any subsequent follow-up questions (M15–M22). Accordingly, we used general editing rules rather than imputation for households where respondents did not know whether there were any additional phone numbers.

A hot deck imputation method involves replacing the missing value from a person who failed to answer a question with a valid value from a “similar” respondent. To decide which criteria should be used in determining a similar respondent, we performed a series of logistical regressions using question M14 (which asks about any additional telephone numbers, excluding cell phones) as the dependant variable. Since most of the households with missing phone information resulted from break-offs, the choice of independent variables was limited to data collected before section L or sampling frame data. Therefore, we were unable to use any demographic information captured in section O of the NSAF survey.

FIGURE 1: NSAF PHONE QUESTIONS

1997 and 1999 NSAF Questionnaire

➢M14. Besides (RESPONDENT’S PHONE NUMBER), do you have other telephone numbers in your household?

▪YES........1 (GO TO M15) ▪NO......... 2 (GO TO NEXT SECTION)

➢M15. How many of these additional telephone numbers are for home use?

▪ NUMBER ____ (GO TO NEXT SECTION)

2002 NSAF Questionnaire

➢M14. Besides (RESPONDENT’S PHONE NUMBER), do you have other telephone numbers in your household, not including cell phones?

▪YES........1 (GO TO M15) ▪NO......... 2 (GO TO NEXT SECTION)

➢M15. Including computer and fax phone numbers, how many of these additional phone numbers are for home use?

▪ NUMBER ____ (IF M15 = 0, GO TO NEXT SECTION, M15 = 1 GO TO M16, M15 > 1 GO TO M17)

➢M16. Is thus additional phone number used for a computer or fax machine?

▪YES........1 (GO TO M20) ▪NO....….. 2 (GO TO NEXT SECTION)

➢M17. Of these (number of phone numbers) additional home use phone numbers, how many are used for a computer or fax machine?

▪ NUMBER ____ (IF M17 = 0, GO TO NEXT SECTION, M17 =1 GO TO M20, M17 > 1 GO TO M19)

➢M19. How many of these (number of phone numbers) phone numbers used for computers or faxes are ever answered for talking?

▪ NUMBER ____ (IF M19 = 0, GO TO NEXT SECTION, M19 =1 GO TO M21, M19 > 1 GO TO M22)

➢M20. Is it ever answered for talking?

▪YES........1 (GO TO M21) ▪NO......... 2 (GO TO NEXT SECTION)

➢M21. Is this phone number used for a computer or fax line answered for:

- personal calls 1 (GO TO NEXT SECTION)
- business calls, or 2 (GO TO NEXT SECTION)
- both? 3 (GO TO NEXT SECTION)

➢M22. Of these (number of phone numbers that are answered, how many are answered for non-business related calls? ▪ NUMBER ____ (GO TO NEXT SECTION)

Using the results of our logistical regression analysis we developed the following set of matching keys:

Site: (14 categories) Alabama, California, Colorado, Florida, Massachusetts, Michigan, Minnesota, Mississippi, New Jersey, New York, Texas, Washington, Wisconsin, Balance of U.S.
Region: (4 categories) Northeast, South, Midwest and West
Screener Income: (3 categories) above 200 % of poverty, below 200% of poverty, or don't know
Total Adults: (2 categories) more than 2 adults, or 2 or fewer adults
Phone Use: (2 categories) sampled phone number also used for business, or not used for business
Language: (2 categories) screener conducted in English or Spanish

We started the imputation process for those respondents who refused or were not asked the first question in the sequence (question M14). Respondents who answered all the phone questions were candidates for providing data for respondents who had missing data. We call these respondents who have valid data "donors". First, donors were matched to respondents with missing data based on the six matching keys listed above. Next, a donor was randomly selected from the "pool" of donors matched at the highest level. The selected donor's value for M14 and the remaining phone questions (M15 to M22) were assigned to the respondent with missing data.

The donors were sorted first by how many of the above five matching keys had the same value as the respondent who had missing data and then sorted randomly. Once a respondent was used as a donor for a particular question, he or she was removed from the donor pool for that question. Therefore, no respondent was used as a donor for the same question more than once.

Except for one important difference the same imputation procedures were used to impute the data for the 39 households who had valid data for M14 but missing data elsewhere in the phone question sequence (M15 through M22). For these cases an additional requirement was made that the selected donor and the donee have the same answer to the question before the question being imputed.

For the 479 households that had missing data on the telephone questions, 13 were deterministically coded, 462 matched on all six matching keys, 2 matched on five of the matching keys (all but the language variables), and 2 matched on four variables (all except the language and phone use variable).

Analytic Concerns and Implications

The length and complexity of the NSAF questionnaire contributed to higher levels of item nonresponse in 2002 than in previous rounds and to the challenges faced in post–data collection editing.

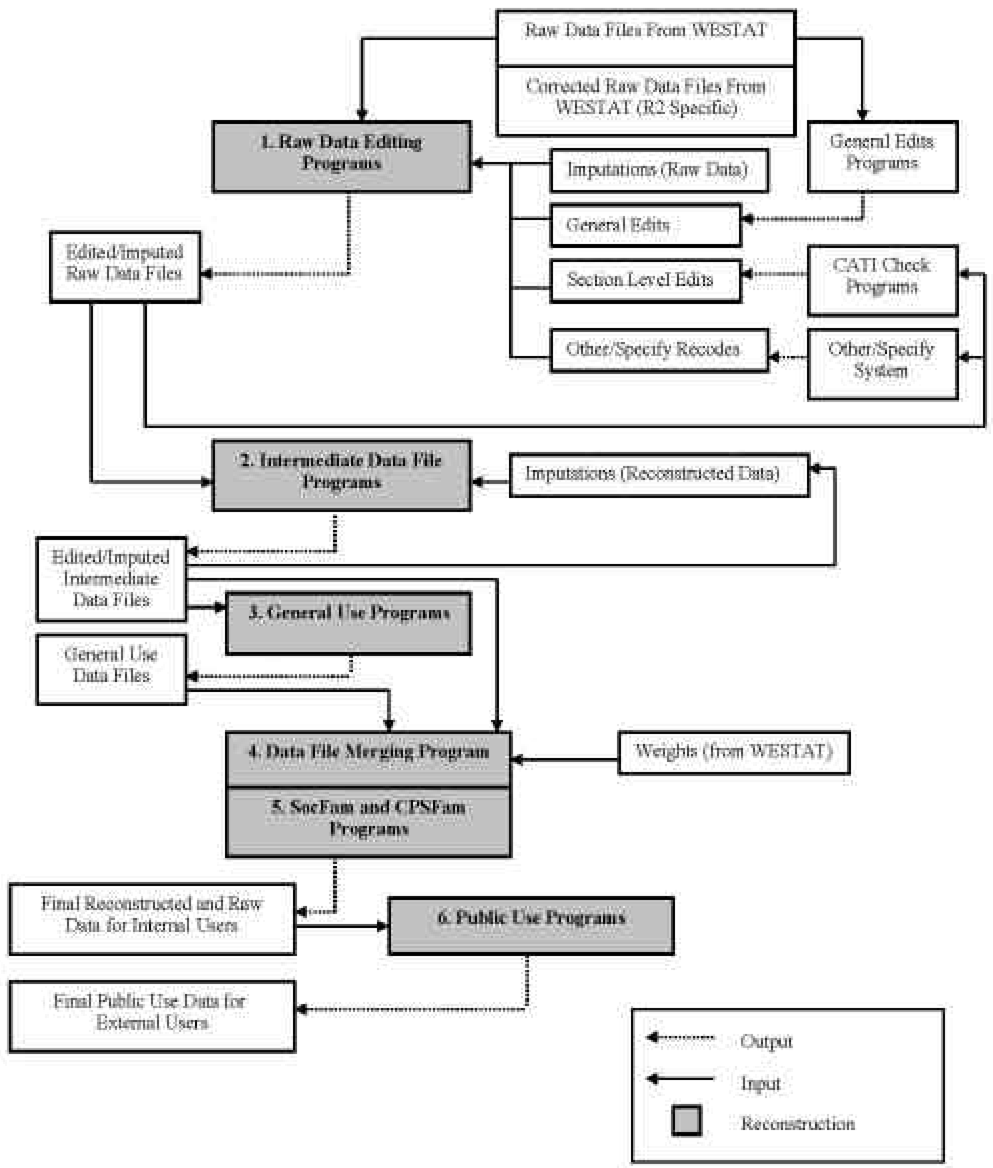
Non-interviews have been extensively covered in Reports No. 7 and 8 in both the 1997 and 2002 Methodology series, and Report No. 8 in the 1999 series. Despite the fact that the amount of unit nonresponse was sizable and raised the cost of the survey, there is little evidence of any serious overall bias after adjustment.

Despite the high quality of the data editing and imputation in the NSAF, researchers still need to be concerned with how they handle the resulting data. Hot deck imputation, for example, can increase the sampling error (Scheuren 1976) in ways that are hard to measure without special procedures, such as multiple imputations (Rubin 1987). In many cases, with low nonresponse, the understatement of variance may be ignored; but for some items, notably income amounts, nonresponse is serious enough to attempt a crude adjustment. One conservative approach for adjusting variance estimations is described in section 4.6 of 1999 NSAF Methodology Report No. 10. There are also some creative approaches (Brick et al. 2004; Kim 2001) that use statistical modeling techniques to estimate the variance associated with items including values resulting from hot deck imputation.

Misclassification concerns arise in any editing or imputation procedure unless the method of assignment perfectly places each missing or misreported case into the right group. The final analyst might employ a re-weighting (or re-imputation) option rather than using the imputations provided. To support this option on the NSAF Public Use Files, we have provided a great deal of diagnostic information, including the imputation flags, replicate weights, sampling variables, and some variables associated with the interview process itself.

As with any data set, researchers will need to be aware of possible anomalies. We believe, however, that these are rare in the NSAF analytical data files. Still, it is unlikely that we have been able to anticipate all the ways the data will be used. Almost certainly errors will be found when researchers carry out their detailed investigations. We would greatly appreciate being informed of any such discrepancies, so they can be brought to the attention of others. Furthermore, depending on the nature of this information, new data sets may be made available.

NSAF DATA DEVELOPMENT OVERVIEW


Raw Data Files From WESTAT
Corrected Raw Data Files From WESTAT (R2 Specific)
1. Raw Data Editing Programs
General Edits Programs
Imputations (Raw Data)
General Edits
Edited/Imputed Raw Data Files
Section Level Edits
CATI Check Programs
Other/Specify Recodes
Other/Specify System
2. Intermediate Data File Programs
Imputations (Reconstructed Data)
Edited/Imputed Intermediate Data Files
3. General Use Programs
General Use Data Files
4. Data File Merging Program
Weights (from WESTAT)
5. SocFam and CPSFam Programs
Final Reconstructed and Raw Data for Internal Users
6. Public Use Programs
Final Public Use Data for External Users
Output
Input
Reconstruction

This flowchart was developed by David D'Orio of the Urban Institute's Information Technology Center.

References

Abi-Habib, Natalie, Adam Safir, and Timothy Triplett. 2004. *2002 NSAF Survey Methods and Data Reliability*. Methodology Report No. 1. Washington DC: The Urban Institute.

Brick, Michael J., Ismael Flores Cervantes, and David Cantor. 1998. *1997 NSAF Response Rates and Methods Evaluation*. Methodology Report No. 8. Washington, DC: The Urban Institute.

Brick, Michael J., et al. 2000. *1999 NSAF Response Rates and Methods Evaluation*. Methodology Report No. 8. Washington, DC: The Urban Institute.

Brick, Michael J., et al. 2003. *2002 NSAF Response Rates*. Methodology Report No. 8. Washington, DC: The Urban Institute.

Brick, Michael J., Graham Kalton, and J.K. Kim. 2004. "Variance Estimation with Hot Deck Imputation Using a Model." *Survey Methodology* 30 (1): 57–66.

Dean Brick, Pat, Genevieve Kenney, Robin McCullough-Harlin, Shruti Rajan, Fritz Scheuren, Kevin Wang, J. Michael Brick, and Pat Cunningham. 1999. *1997 NSAF Survey Methods and Data Reliability*. Methodology Report No. 1. Washington, DC: The Urban Institute.

Dipko, Sarah, Michael Skinner, Nancy Vaden-Kiernan, John Coder, Esther Engstrom, Shruti Rajan, and Fritz Scheuren. 1999. *1997 NSAF Data Editing and Imputation*. Methodology Report No. 10. Washington, DC: The Urban Institute.

Dipko, Sarah, Michael Skinner, Nancy Vaden-Kiernan, Tamara Black, John Coder, Nathan Converse, Veronica Cox, Aparna Lhila, and Fritz Scheuren. 2000. *1999 NSAF Data Editing and Imputation*. Methodology Report No. 10. Washington, DC: The Urban Institute.

Ford, B. 1983. "An Overview of Hot-Deck Procedures." In *Incomplete Data in Sample Surveys, Vol. 2*. Burlington, MA: Academic Press, Inc.

Groves, Robert M., and Douglas Wissoker. 1999. *1997 NSAF Early Nonresponse Studies*. Methodology Report No. 7. Washington, DC: The Urban Institute.

Kim, J.K. 2001. "Variance Estimation with Hot Deck Imputation Using a Model." *Survey Methodology* 27(1): 75–83.

Rubin, D. 1977. "Formalizing Subjective Notions about the Effect of Nonrespondents in Sample Surveys." *Journal of the American Statistical Association* 72: 538–43.

Rubin, D. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

Scheuren, F. 1976. "Preliminary Notes on the Partially Missing Data Problem—Some (Very Elementary) Considerations." Paper delivered at the April meeting of the Social Security Administration's Statistical Methodology Group, 1976.