

2002 NSAF Collection of Papers

Report No. 6

Prepared by:

Natalie Abi-Habib
Tamara Black
Simon Pratt
Adam Safir
Rebecca Steinbach
Timothy Triplett
Kevin Wang
The Urban Institute

J. Michael Brick
David Cantor
Patricia Cunningham
David Ferraro
Benmei Liu
David Martin
Patricia Warren
Erin Wilson
Westat

John Wivagg
DataSource



Assessing
the New
Federalism

An Urban Institute
Program to Assess
Changing Social Policies

Methodology Reports

PREFACE

2002 NSAF Collection of Papers is the sixth 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 over 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. 6 is a collection of occasional papers on technical issues in the design, implementation, and operation of the 2002 round of the NSAF. It is a companion report to the 1999 methodology series *Report No. 7 NSAF Collection of Papers* and the 1997 methodology series *Report No. 16 NSAF Technical Papers*. All the papers in this collection were presented at either the annual May American Association for Public Opinion Research conference or the annual August Joint Statistical Meetings.

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

<u>Chapter</u>		<u>Page</u>
1	PAPERS PRESENTED AT THE AMERICAN ASSOCIATION FOR PUBLIC OPINION RESEARCH CONFERENCE	1-1
	Overview.....	1-1
	Effects on Survey Estimates from Reducing Nonresponse	1-12
	Determining the Probability of Selection for a Telephone Household in a Random Digit Dial Sample Design Is Becoming More Difficult.....	1-11
	An Experiment in Call Scheduling for an RDD Survey	1-18
	Increased Efforts in RDD Surveys.....	1-28
	Comparing Incentives at Initial and Refusal Conversion Stages on a Screening Interview for a Random Digit Dial Survey	1-36
	Comparing Promised and Prepaid Incentives for an Extended Interview on a Random Digit Dial Survey	1-46
2	PAPERS PRESENTED AT THE AMERICAN STATISTICAL ASSOCIATION'S ANNUAL JOINT STATISTICAL MEETINGS....	2-1
	Overview.....	2-1
	Using Paradata to Examine the Effects of Interviewer Characteristics on Survey Response and Data Quality.....	2-2
	Using a Short Follow-up Survey to Compare Respondents and Nonrespondents	2-11
	Sampling Refusals: Why, When, and How Much?	2-19
	Trimming Extreme Weights in Household Surveys	2-24

1. PAPERS PRESENTED AT THE AMERICAN ASSOCIATION FOR PUBLIC OPINION RESEARCH CONFERENCE

Overview

Chapter 1 consists of six NSAF methodology papers that were presented at the American Association for Public Opinion Research (AAPOR) conference and have not been published elsewhere besides in the conference proceedings. While most of these papers can be found in the AAPOR proceedings, the versions included in this report may vary because of the page size and time deadlines associated with papers submitted for proceedings publication.

All papers in this chapter are based on the 2002 NSAF data collection experience, except for the first paper (“Effects on Survey Estimates from Reducing Nonresponse”). The first paper uses 1999 data and is included in this report because it was presented after the release of the 1999 collection of papers methodology report.

Effects on Survey Estimates from Reducing Nonresponse

Adam Safir, Rebecca Steinbach, Timothy Triplett, and Kevin Wang

1. Introduction

Using a variety of procedures designed to maximize response rates, survey organizations expend sometimes extraordinary efforts to minimize the potential for nonresponse bias. Given that nonresponse bias is a function of both the nonresponse rate and the difference between respondents and nonrespondents, maximizing response rates is a sensible approach to minimizing the potential for bias contributed by less-than-perfect survey participation rates.

The second part of the equation is by definition a more complicated component to address. A number of approaches have been suggested for measuring the size of the difference between respondents and nonrespondents, despite the unobserved status of the latter. One such method is to use difficult-to-interview respondents, obtained through increased call attempts, higher incentives, or an extended field period, as proxies for nonrespondents.

Although it is assumed that additional efforts to obtain interviews with the difficult-to-interview will improve precision and reduce nonresponse bias (Lynn et al. 2002), when the interviews obtained as a result of these efforts display characteristics similar to interviews already conducted with easier-to-interview respondents, researchers may arrive at one of two conclusions: (1) the difficult-to-interview, or nonrespondents-by-proxy, do not differ in meaningful or systematic ways from other respondents, thus implying ignorable nonresponse; or (2) a core group of nonrespondents remains unmeasured, thus suggesting the potential for nonignorable nonresponse bias. Faced with either prospect, researchers may question the extent to which additional interviewing efforts are merited, given the absence of identifiable nonignorable nonresponse bias (1) or apparent ineffectiveness (2). For example, where there is little indication of a bias reduction resulting from extended efforts to obtain additional interviews, the survey organization may consider a redesign of expensive refusal reworking procedures (Scheuren 2000).

This paper presents the results of research conducted to analyze the effects of efforts to minimize the potential for nonresponse bias in the 1999 round of the National Survey of America's Families (NSAF). In particular, this research was motivated by questions about the efficacy of maximizing response rates on minimizing nonresponse bias.

In the first major analysis component—level of effort—we address the effect that increasing the level of effort expended to increase participation rates has on reducing nonresponse bias by comparing the characteristics of persons in easy-to-interview households to the characteristics of persons in difficult-to-interview households. These groups are defined by number of calls to contact and number of refusals. As the literature has suggested that the characteristics of noncontacts and refusals may differ substantively from each other as well as from the “average” respondent, particular emphasis was given to examining differences between each subset of the difficult-to-interview, the difficult-to-contact (5+ calls to contact) and the reluctant-to-participate (2+ refusals), and the average interviewed household. In addition to comparing these groups

within the 1999 survey round, we compare measures associated with varying levels of contactability and cooperation across survey rounds.

The second major analysis component—potential for nonresponse bias—focuses on assessing the potential for nonresponse bias due to unmeasured sample elements, treating difficult-to-interview observations as informative of the noninterviewed. Within this analysis step, we also report on the results of a comparison of sampling frame data across easy-to-interview, difficult-to-interview, and noninterviewed households, defined by completion status in NSAF and a short follow-up survey. The data were compared across these three groups to assess the appropriateness of using difficult-to-interview respondents as proxies for the noninterviewed.

2 Data Sources

This research uses data from the 1999 and 1997 rounds of NSAF data collection, as well as data from a nonresponse follow-up survey to the 1999 NSAF survey. The NSAF is a survey of the economic, health, and social characteristics of children, adults under the age of 65, and their families. The survey has a dual-frame design (random-digit-dial of telephone households and area sample of nontelephone households), features an oversample of low-income households with children, and is representative of the nation and of 13 states. The questionnaire consists of a short screening interview, used to determine household eligibility, and a longer extended interview, used to gather detailed information on the characteristics of sampled household members. On average, the interview lasts 30 to 45 minutes, and is conducted with the most knowledgeable adult (MKA) of the sampled child/ren and/or a randomly sampled childless adult in a subset of households. The NSAF uses standard survey methods to reduce nonresponse, such as multiple contact attempts and refusal conversion, as well as more extensive efforts, including monetary incentives and an extended field period. Westat conducted both rounds of data collection for the NSAF.

The second data source consisted of a short follow-up survey (SFS) conducted with a random selection within predefined strata of NSAF respondents and nonrespondents. The starting sample size was 2,000 finalized NSAF telephone numbers, of which 1,788 were determined eligible for interview or reinterview. The questionnaire included selected items from the NSAF instrument, as well as opinion questions about the importance of surveys and research. The data collection for SFS was conducted by the University of Maryland Survey Research Center (SRC) during the later stages of the 1999 NSAF field period.

3 Prior Research

A number of nonresponse studies were conducted following the first round of NSAF data collection (1997) to learn more about the characteristics of NSAF nonrespondents and to assess the impact of missing data from unit nonresponse on survey estimates. We apply the same basic approach of these earlier nonresponse analyses to the 1999 NSAF data, and compare survey results among respondents by level of effort required to obtain an interview, with the assumption that the results of these comparisons would be informative of the differences between those interviewed and those not interviewed.

In the 1997 nonresponse analyses, it was expected that “any pattern for the socioeconomic indicators would be consistent with two hypotheses about the influences toward participation in the NSAF—that those receiving transfer payments would be at home more often (and thus more easily contacted and perhaps with lower time costs of participation) and, because of the topic of the survey, that those receiving transfer payments would be more interested in providing information to the interviewer. Both of these observations are important because they suggest the possibility of nonignorable nonresponse errors; that is, both for contact and for cooperation, the attribute of key interest is an indirect causal factor for response” (Groves and Wissoker 1999).

For statistics computed on 1997 NSAF telephone households with children, little evidence of important nonresponse errors was observed. However, there was a small tendency for households with higher socioeconomic status to require more effort to obtain an interview. Additionally, NSAF nonrespondents tended to be black non-Hispanic (Groves and Wissoker 1999). Overall, no evidence for a serious nonresponse bias arising from a large fraction of refusals was detected.

Other studies have found that reluctant respondents tended to be older, with somewhat lower socioeconomic status, while difficult-to-contact respondents tended to be younger and more affluent (Curtin et al. 2000).

4 Level of Effort

4.1 Methods

The level of effort analysis examined the effect on estimates from reducing nonresponse. As previously noted, this research was motivated by an interest in understanding the gains in nonresponse bias reduction realized from the additional effort expended to obtain interviews with the difficult-to-interview. Most large-scale data collection efforts with limited resources face the same need to address whether level of effort should be increased to improve data quality, albeit at a higher cost, or whether it can be reduced to minimize operational costs, without a corresponding risk to data quality.

The level of effort analysis file was restricted to 1999 NSAF children in telephone households. The total sample size was 34,831 sampled children. Comparison groups were formed by classifying the sample into hierarchies of contactability based on number of calls before first screener contact (1, 2, 3 or 4, and 5+), and cooperation, based on number of refusals before completing the interview (0, 1, and 2+). We included both screener and extended refusals in the total number of refusals; while there may be some differences between the two types, our primary interest was in the presence of a reluctant household member, regardless of the type of reluctance (initial refusal versus second interview refusal).

Table 1. Estimates by Number of Calls Required for First Contact, RDD Cases with Children

Characteristic	Calls Before First Screener Contact				Total
	One	Two	3 or 4	5+	
Foreign-born person lives in household	16.22	13.41	13.38	12.51	15.06
s.e.	0.47	0.79	1.09	1.15	0.39
Household income below 200% poverty level	35.26	33.66	32.90	26.78	33.99
s.e.	0.75	1.06	1.30	1.46	0.53
Received food stamps in 1998	10.76	10.11	8.17	9.79	10.23
s.e.	0.46	0.57	0.75	1.32	0.34
Both MKA and spouse employed	60.87	61.09	63.32	69.92	61.87
s.e.	0.67	1.57	1.76	2.03	0.54
No high school degree or GED (MKA)	10.70	9.50	7.67	5.59	9.70
s.e.	0.43	0.67	0.87	0.79	0.31
Hispanic (MKA)	13.68	13.24	11.39	9.44	12.98
s.e.	0.41	0.80	0.94	1.01	0.34
Age of MKA	37.57	37.54	36.82	36.84	37.41
s.e.	0.09	0.19	0.21	0.22	0.07
MKA has health insurance	84.81	86.85	87.08	89.26	85.81
s.e.	0.44	0.77	0.99	1.09	0.31

The analysis compared the household and parental characteristics of children in difficult-to-contact and reluctant-to-participate households with those of children in the “average” responding household, using selection weights. The selection weights excluded any adjustments for nonresponse or poststratification. We included a broad range of measures, intended to replicate the 1997 analyses by Groves and Wissoker, and to reflect a variety of demographics and other survey items, such as age, race, ethnicity, education, income, employment, health insurance, program participation, family structure, and household tenure. We expected the findings to inform an assessment of the relative utility of additional contact and conversion attempts, although a final determination is limited by the lack of revised weights to compare estimates with and without the cases obtained as a result of additional efforts.

In addition to examining potential differences in these groups within the 1999 data, we compare group differences across survey rounds. Using the results from the Groves and Wissoker analysis, we compare the differences between the groups in 1997 with the differences between the groups in 1999, looking at the difference of the differences to gauge change in degree or direction.

4.2 Results

The additional effort expended to interview difficult-to-contact and reluctant-to-participate respondents yielded respondents whose characteristics and circumstances were relatively similar to those of the more easily interviewed, with some notable differences.

Table 1 compares attributes of children in households easily contacted with the attributes of those in difficult-to-contact households. Parents in the difficult-to-contact households tend to be less poor, less Hispanic, more insured, younger, and less foreign-born. Further, parents of children in easily contacted households were more likely to be unemployed or not in the labor force, and more likely to receive benefits from the government. Parents in the more difficult-to-contact households tended toward higher socioeconomic status and education levels. While

intuitive and consistent with prior research, these results do reemphasize the importance of additional contact attempts, particularly as children living with employed adults tended to have been contacted only with a greater number of calls.

Table 2 presents a comparison of children in households interviewed without a refusal, those interviewed after one refusal, and those interviewed after two or more refusals.

Table 2. Estimates by Refusal Status, RDD Cases with Children

Characteristic	Refusal Status			Total
	None	One	2+	
Foreign-born person lives in household	16.07	13.75	12.40	15.06
s.e.	0.45	0.66	0.83	0.39
Homeowner lives in household	70.76	75.64	76.95	72.72
s.e.	0.59	0.75	1.25	0.36
Household income below 200% poverty level	35.61	31.19	31.46	33.99
s.e.	0.65	0.93	1.37	0.53
Received food stamps in 1998	10.88	8.77	9.99	10.23
s.e.	0.42	0.64	1.15	0.34
Confident medical care available if needed	92.99	93.76	93.01	93.2
s.e.	0.36	0.49	0.66	0.3
Unable to pay mortgage, rent, or utilities, past year	15.41	14.03	14.06	14.9
s.e.	0.53	0.76	1.03	0.42
Ever skip meals because money unavailable	10.77	10.49	9.88	10.6
s.e.	0.44	0.55	0.92	0.36
Black, non-Hispanic (MKA)	10.45	10.51	15.55	11.01
s.e.	0.40	0.63	1.41	0.34
Age (MKA)	37.12	37.79	38.05	37.41
s.e.	0.10	0.13	0.22	0.07

Reluctant-to-participate households tend to be more likely to include homeowners, be slightly less poor, more black non-Hispanic, older, and less foreign-born. Larger, significant differences were seen in some measures, particularly on the demographic items such as race and age. However, it is important to note that these observed, larger demographic/socioeconomic differences did not translate into significant differences in other important outcome items, such as access to care, food security, or economic hardship.

With respect to changes in the degree and direction of differences over time, table 3 compares the difference between groups in the 1997 data with the difference between groups in the 1999 data. For example, in 1997, the percentage point difference between past year TANF receipt in 2+ refusal households and the average was -0.6 (6.1 versus 6.7), or -9.0 percent. In 1999, this difference shrank to -0.4 percentage points (4.2 versus 4.6) or -8.1 percent,¹ resulting in a 0.2 point decrease and 0.9 percentage point decrease (-9.0 percent versus -8.1 percent) in the difference of these groups from 1997 to 1999.

¹ As a result of rounding, some differences may appear to be slightly higher or lower than the difference of the reported rates.

Overall, the comparisons in table 3 show that the degree of difference between difficult-to-interview and average households has decreased, with one or two exceptions, such as race and ethnicity. It should be noted that although there were some differences between the incentive structure used for refusal conversion in round 1 and the incentive structure used in round 2, it is our belief that the increased use of incentives in the second round would have served to increase the estimated difference between the difficult-to-interview and the average household (due to higher conversion rates among the more reluctant sampled households). In fact, the opposite was observed.

We speculate that the decrease in the degree of difference may be attributable to a number of factors. For example, owing to social or telephony changes, the “pool” of the harder-to-interview may be increasing to include those whose characteristics and circumstances are more similar to the easier-to-interview, leading to a capture of a greater number of difficult-to-interview households that more closely resemble easier-to-interview cases (and are therefore potentially less informative of nonrespondents), and proportionately fewer difficult-to-interview households that are more similar to nonrespondents (and who are arguably more informative of nonrespondents).

Table 3. Degree and Direction of Differences in Survey Estimates by Level of Effort and Survey Round

	Round 1		Round 2		R2 Diff vs. R1 Diff	
	Pt. Diff	% Diff	Pt. Diff	% Diff	Pt. Diff	% Pt. Diff
5+ Calls-to-Contact vs. Average						
Homeowner lives in household	1.1	1.5	0.0	0.0	-1.1	-1.5
Household income below 200% poverty level	-8.7	-24.6	-7.2	-21.2	-1.5	-3.4
Received TANF last year	-2.5	-37.3	0.8	16.8	-1.7 *	-20.5 *
Confident medical care available if needed	2.2	2.4	0.6	0.6	-1.6	-1.8
Ever skip meals because money unavailable	-1.5	-12.5	-1.0	-9.2	-0.5	-3.3
Biological mother lives in household	-1.8	-2.0	-0.1	-0.2	-1.7	-1.8
2+ Refusals vs. Average						
Household income below 200% poverty level	-3.5	-9.9	-2.5	-7.4	-1.0	-2.4
Received TANF last year	-0.6	-9.0	-0.4	-8.1	-0.2	-0.9
Confident medical care available if needed	0.3	0.3	-0.2	-0.2	-0.1 *	-0.1 *
Ever skip meals because money unavailable	-1.5	-12.5	-0.7	-6.8	-0.8	-5.7
Black, non-Hispanic (MKA)	0.2	1.9	4.5	41.2	4.3	39.3
Hispanic (MKA)	0.5	4.4	-3.3	-25.7	2.8 *	21.2 *

* Indicates significant differences at the .05 level.

5. Potential for Nonresponse Bias

5.1 Methods

The appropriateness of using the difficult-to-interview as a proxy for the noninterviewed rests on the validity of the assumption that the difficult-to-interview characteristically resemble the noninterviewed. To test this assumption, we use exchange-level sampling frame data to examine differences between households that completed the NSAF (Group AB), households that completed the SFS but not the NSAF (Group C), and households that did not complete either the NSAF or the SFS (Group D).

The exchange-level data were provided on the Genesys Sampling Systems sample data file. As projections based on FIPS county projections for dominant exchanges, the exchange-level data have a certain level of coarseness, but are still useful data to analyze, particularly since they are used to form the nonresponse weighting adjustment classes in NSAF. Additionally, the exchange-level data feature a desirable level of geographic specificity.

**Exhibit 1. Sampling Frame Data
Comparison Groups**

Interview Status		SFS	
		Yes	No
NSAF	Yes	Group A (n = 675)	Group B (n = 318)
	No	Group C (n = 231)	Group D (n = 562)

T-tests were used to compare the mean characteristics of Group AB (the easy-to-interview) with Group C (the difficult-to-interview) and Group D (the noninterviewed), as well as to compare the mean characteristics between Groups C and D. Under the assumption that the difficult-to-interview are informative of the noninterviewed, we expected to see small or no differences between Groups C and D, and larger differences between either or both of these two groups and the easy-to-interview group, Group AB. We include “B” in the easy-to-interview group because although interviews were attempted but not obtained in SFS, we acknowledge that the length of the NSAF interview likely had an effect on the decision to participate in the follow-up (for respondents who completed the NSAF but not the SFS).

Comparison measures included average rent, median income, median home value, percent age 0–17, percent black non-Hispanic, percent Hispanic, percent renters, percent listed, percent income \$0–\$10K, percent income \$11–\$15K, and percent income \$16–\$25K. The results of the sampling frame data would be used to draw conclusions about the appropriateness of making statements about the potential for nonresponse bias due to unobserved sample elements.

5.2 Results

At the exchange-level, Group C (difficult-to-interview) respondents tended to live in exchanges with a higher percentage of black non-Hispanics and renters and a lower percentage of listed telephone numbers than Group AB respondents (easier-to-interview). Alternatively, Group D households (noninterviewed) were shown to live in exchanges with a significantly higher median income, higher average rent, higher percent black non-Hispanic, higher percent Hispanic, and lower percentage of listed telephone numbers than did Group AB respondents (see table 4).

Table 4. Exchange-level Household Estimates by Comparison Group

Exchange Characteristic	Group		
	AB	C	D
Average Monthly Rent (in dollars)	464	476	503
Median Income (in dollars)	42,002	40,919	44,528
Percent Black	11.4	14.8	12.1
Percent Hispanic	8.3	9.6	10.0
Percent Listed Telephone Numbers	38.6	35.9	36.9
Percent Renters	33.7	37.4	34.8
Percent Age 0-17	25.7	25.7	25.2

Overall, Group AB households are more similar to Group D households and less similar to Group C households; however, Group C and D households exhibit smaller between-group differences, and both exhibit larger differences compared with Group AB households. The implication, based on the available data, is that Group C households, and respondents living in such households, may be viewed as reasonable proxies for respondents living in Group D households. However, the differences between the three groups are almost negligible across most measures. Given such small differences, the utility of the exchange-level data may be limited in arriving at actionable conclusions.

6. Conclusions

The results of our research indicate that on average, the characteristics of children in difficult-to-contact and reluctant-to-participate households do not differ in meaningful ways from those of children in average households. Although larger differences were seen the demographic makeup of the groups (for example, in education and employment between contactability groups, and in race and ethnicity between cooperation groups), these differences were not observed to carry over into important outcome measures, such as confidence in medical care and food insecurity. While these results were encouraging, we acknowledge certain analytic limitations, such as the lack of revised selection weights to compare estimates with and without the difficult-to-interview cases.

We also note a perceptible decrease in the degree of difference between these groups over time, although again, we lack sufficient data to draw substantive conclusions from this finding. We speculate that as the pool of more difficult-to-interview households grows “passively” due to the increased availability and use of telephony barriers (e.g., caller ID, dual voice-computer lines), “new” difficult-to-interview households may exhibit fewer differences than easier-to-interview households. While this may suggest that some increase in absolute nonresponse may not translate to a monotonic increase in potential for bias, it does raise the specter of a core group of difficult-to-interview households that have become even more difficult to identify and interview within a now larger difficult-to-interview respondent pool. Additionally, the increase of difficult-to-interview households that now more characteristically resemble easier-to-interview cases may

further undermine the assumption that the difficult-to-interview are informative of the noninterviewed, thereby diminishing their utility as proxies for the non-interviewed.

With respect to using the difficult-to-interview as proxies for the noninterviewed, the analysis of sampling frame data showed that difficult-to-interview and noninterviewed households were more similar at the exchange level, and each less similar to the easier-to-interview. While this points positively to the use of the difficult-to-interview as proxies for the noninterviewed, the coarseness of the sampling frame data limit our ability to examine these findings in more detail.

References

Curtin, Richard, Stanley Presser, and Eleanor Singer. 2000. "The Effects of Response Rate Changes on the Index of Consumer Sentiment." *Public Opinion Quarterly* 64(4).

Groves, Robert, and Doug Wissoker. 1999. *Early Nonresponse Studies of the 1997 National Survey of America's Families*. Methodology Report No. 7. <http://www.urban.org>.

Keeter, Scott, Carolyn Miller, Andrew Kohut, Robert M. Groves, and Stanley Presser. 2000. "Consequences of Reducing Nonresponse in a National Telephone Survey." *Public Opinion Quarterly* 64(2).

Lynn, Peter, P. Clarke, J. Martin, and P. Sturgis. 2002. "The Effects of Extended Interviewer Efforts on Nonresponse Bias." In *Survey Nonresponse*, edited by R.M. Groves, D.A. Dillman, J. Eltinge, and R.J.A. Little (135–47). New York: John Wiley.

Scheuren, Fritz. "Quality Assessment of Quality Assessment." *American Statistical Association 2000 Proceedings of the Section on Government Statistics and Section on Social Statistics*. Alexandria, VA: American Statistical Association.

Determining the Probability of Selection for a Telephone Household in a Random Digit Dial Sample Design Is Becoming More Difficult

Timothy Triplett and Natalie Abi-Habib

1. Introduction

Over the past 25 years, a lot of household telephone surveys have made use of a random digit dial (RDD) sample design (Waksberg 1978). During the early years of RDD sampling the vast majority of telephone households had only one telephone number. By 1988, just 2.7 percent of telephone households had more than one noncellular telephone number.² This figure steadily increased and by the year 2000, 26.2 percent of telephone households had more than one telephone number. Accounting for the rapid increase in multiple telephone line households are advancements in telecommunication technologies such as the Internet and fax machines, as well as an increase in home businesses. However, since 2000 the trend has reversed, causing a dip in the percentage of telephone households having more than one noncellular phone number (or landline number). The drop may well be attributed to further telecommunication technology advancements whereby a single phone line can serve the multiple purposes of voice communication and Internet or fax machine connection simultaneously. Also contributing to the decline in multiple telephone number households may be the recent surge in cellular telephone purchases to replace additional household landlines.

Recent rapid changes in the telecommunications culture have created a challenge for survey researchers attempting to measure the number of household telephone numbers eligible for selection into a sample. Respondents may be unsure about what counts as a household telephone number if researchers fail to be specific in their definitions.

Today, more than a fifth of the U.S. household telephone population maintains two or more telephone numbers. A proportion this large demands that survey researchers obtain reliable data about how many telephone numbers a household has that are part of the RDD sample frame and that are always or sometimes used for personal (non-business) conversations.

For many years, researchers using a RDD sample design could estimate the total number of residential telephone numbers in a household by simply asking one, sometimes two, and at most three questions. The 2002 National Survey of America's Families (NSAF) is a telephone survey that relies primarily on a large RDD sample design using over 400,000 telephone numbers. In previous rounds of the NSAF (1999 and 1997) a simple two-question approach was used to estimate a household's total number of sample eligible residential telephone numbers. For the 2002 study, a more in-depth set of questions was asked of each household, with the purpose of learning whether additional telephone numbers could be used for completing a survey. This paper compares the results of these questions with previous rounds of NSAF and looks at what other RDD studies are doing.

² According to the Federal Communications Commission's August 2003 report, "Trends in Telephone Service."

Figure 1. NSAF Telephone Question Series 1997, 1999 and 2002

1997 and 1999 NSAF Questionnaire

M14. Besides [TELEPHONE NUMBER], do you have other telephone numbers in your household?

Yes \longrightarrow [Go to M15] No \longrightarrow [Go to next section]

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

NUMBER \longrightarrow [Go to next section]

2002 NSAF Questionnaire

M14. Besides [TELEPHONE NUMBER], do you have other telephone numbers in your household, not including cell phones?

Yes \longrightarrow [Go to M15] No \longrightarrow [Go to next section]
 DK \longrightarrow [Go to M18]

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

NUMBER M15 = 0 \longrightarrow [Go to next section]
 M15 = 1 \longrightarrow [Go to M16]
 M15 > 1 \longrightarrow [Go to M17]

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

Yes \longrightarrow [Go to M20]
 No \longrightarrow [Go to next section]

M17. Of these [NUMBER] additional home use phone numbers, how many are used for a computer or fax machine?

NUMBER M17 = 0 \longrightarrow [Go to next section]
 M17 = 1 \longrightarrow [Go to M20]
 M17 > 1 \longrightarrow [Go to 19]

M18. Do you have any additional phone numbers for computer or fax machines?

Yes \longrightarrow [Go to M20]
 No \longrightarrow [Go to next section]

M19. How many of these [NUMBER OF PHONE NUMBERS] phone numbers used for computers or faxes are ever answered for talking?

NUMBER M19 = 0 \longrightarrow [Go to next section]
 M19 = 1 \longrightarrow [Go to M21]
 M19 > 1 \longrightarrow [Go to M22]

M20. Is it ever answered for talking?

Yes \longrightarrow [Go to M21]
 No \longrightarrow [Go to next section]

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

Personal calls
 Business calls
 Both? \longrightarrow [Go to next section]

M22. Of these [NUMBER OF PHONE NUMBERS THAT ARE ANSWERED], how many are answered for non-business related calls?

NUMBER \longrightarrow [Go to next section]

Figure 2. Telephone Question Series from other Major RDD Surveys

California Workforce Study 2001/2002

Tel1. Next, how many telephones do you have in your home—counting extensions, but not counting cellular phones?

Tel2. Do [all/both] the telephones have the same number?

Tel3. How many different numbers are there?

[Tel4 & Tel5: ASK ONLY OF RESPONDENTS WHO HAVE MORE THAN 1 TELEPHONE NUMBER]

Tel4. Are any of those numbers used *exclusively* for computers or fax machines?

IF YES: How many?

Tel5. How many of those lines are used for making or receiving calls for personal or business purposes?

Behavioral Risk Factor Social Survey 2001 and 2002

Q1. Do you have more than one telephone number in your household? Do not include cell phones or numbers that are only used by a computer or fax machine.

Q2. How many of these are residential?

Behavioral Risk Factor Social Survey 1999 and 2000

Q1. Do you have more than one telephone number in your household?

Q2. How many residential phone numbers do you have?

Community Tracking Survey 1998/1999

H30. Do you have any other telephone numbers in you household besides [FILL IN PHONE NUMBER]?

[IF YES]: How many?

H31. (Is this/Are these) other phone numbers for...

Home use

Business and home

Business use

2. Other RDD Surveys

In reviewing what other RDD surveys are doing to determine the probability of a telephone household's selection, it is encouraging to see that most surveys have modified their questions on telephone use in order to address changes in technology. Figures 1 and 2 show the question wording used in all three rounds of the NSAF and in several recent large RDD surveys. What is troubling is that while the information surveys need to obtain is the same, the approaches different surveys are using to get this information are quite varied. Therefore, it is not easy to compare estimates across studies of reported additional residential telephone numbers that require a weighting adjustment.

Survey researchers agree that obtaining correct estimates of residential phone numbers is needed for RDD surveys. Carefully crafted questions with a specific definition for the term "residential telephone number" will aid in accurate measurement. However, as we adopt new procedures it is important that we investigate the effects of new questions on the weighting adjustment associated with multiple residential telephone numbers.

3. NSAF Methodology

Westat collected the data for all three rounds of the NSAF: 1997, 1999 and 2002. The purpose of the NSAF survey is to assess the impact of recent changes in the administration of a number of assistance programs for children and the poor. The sample is based on two different frames, the largest of which is a RDD frame representing households with telephones.³ The second is an area frame from which nontelephone households are selected. All interviews are administered by telephone (interviews in the area frame are conducted through cellular telephones supplied to respondents). The NSAF sample is designed to generalize to 13 specific states, as well as the nation as a whole. The design also includes an oversample of households estimated to be under 200 percent of the federal poverty level as well as households with children.

The NSAF consists of both a screening and an extended interview. The screening interview is designed to assess household eligibility and select a respondent for the extended interview when a household is eligible. Household eligibility for the extended interview is determined by residence of persons less than 65 years of age and by family poverty compared with 200 percent of the federal poverty level. The extended interview is between 30 and 50 minutes in length and covers a wide range of topics, including health, education, child care, income, and receipt of social services.

Questions about the number and use of residential telephone numbers are asked toward the end of the extended interview. Figure 1 shows how the telephone assessment questions were asked on the 1997 and 1999 questionnaire compared with the 2002 version. While the total number of questions and the wording of questions M14 and M15 changed from 1997 and 1999 to 2002, the intent of these questions remained the same: to determine the number of sample eligible telephone numbers in a household. In the 2002 questionnaire, it was considered necessary to explicitly ask respondents *not to include* cellular telephone numbers due to the rapid increase in cell phone usage. The 1997 and 1999 questionnaires included this as an online interviewer instruction that was only read at the respondent's request. Likewise, question M15 of the 2002 NSAF explicitly instructed respondents to include home computer and fax numbers only if they were also used for voice communication. The 1997 and 1999 NSAF surveys again relied on an interviewer instruction. Therefore, while the questions used between rounds of the NSAF were different, the goal of estimating total residential telephone numbers also used for voice communication remained.

In addition to changes in question wording, the 2002 NSAF also includes a series of questions asked of multi-telephone households to assess whether supplemental telephone numbers were sample-eligible. Many people purchase additional telephone numbers for their home computer or fax machine. In some households, telephone numbers used by a computer or fax machine are never answered for voice communication, while in other households the opposite is true. Regardless of their household use, these telephone numbers are included in the sampling frame. We felt the need, on the 2002 NSAF, to ask a set of usage questions to determine whether additional household telephone numbers were available for completing a survey.

³ Over 500,000 phone numbers were selected for the 2002 NSAF.

4. Results from Changing the NSAF Phone Number Items

Since the NSAF telephone questions are asked toward the end of the first extended interview, telephone information was only collected for households that were selected into the study. To analyze the 2002 NSAF telephone question series, we use a household weight adjustment that controls for the probability of a household's selection. This weight adjustment does not include the multiple telephone line adjustment. The weight is an accurate assessment of telephone information for households with a resident 18 to 64 years old. Since we do not collect information from elderly households, where all household members are 65 or older, our data presumably varies somewhat from household surveys that include all telephone households. Table 1 displays estimates of the percent of NSAF multiple telephone households for all three rounds using the adjusted household weight described.

Table 1. NSAF 1997–2002: Percent of Households with Multiple Residential Telephone Numbers (excluding cell phones)

	NSAF	
	1999	2002
1997	14.0%	11.7%
1999	18.3%	

Timeline data from the Federal Communications Commission's August 2003 *Trends in Telephone Service* show increasing numbers of U.S. households acquiring multiple telephone numbers during the 1990s (table 2). These figures exclude cellular telephones but do include telephone numbers purchased only for home business use. The four point increase in the percentage of households with multiple telephone numbers (14.0 percent in 1997 versus 18.3 percent in 1999—table 1) from 1997 to 1999 on the NSAF is consistent with the five point FCC increase. The slightly higher increase in FCC numbers can be explained by possible overreporting of telephone lines due to respondent inclusion of cell phones and/or computer, fax or home business lines not used for personal conversations. In support of this argument we look to the 1999 adult special study supplement to the National Household Education Survey (NHES).⁴ The NHES found that 41 percent of households in 1999 owned at least one cellular telephone. Of these households, 4 percent admittedly included their cellular telephones in their count of additional residential telephone numbers (Roth, Montaquila, and Brick 1999). Likewise, it is probable that some of the 1999 NSAF respondents living in households owning cellular telephones also included them in their count of residential telephone lines. Cellular telephone inclusion likely had less impact on 1997 NSAF telephone line estimates since, according to the Cellular Telecommunications and Internet Association (CTIA), there were 60 percent fewer cellular subscribers in 1997 than in 1999.

Since reaching a peak in 2000, the percentage of households having more than one landline telephone number has dropped, as reported by the FCC. NSAF estimates for the same period show a similar decline: between 1999 and 2002 a 6½ point drop in households with multiple residential telephone numbers is evident (18.3 percent to 11.7 percent—see table 1). Using the

⁴The adult special study supplement to the NHES was conducted by Westat to gather information on telephone technologies that could affect survey response rates or weighting procedures.

NHES 1999 estimates of households owning a cellular telephone and respondents who included cellular telephones in their count of residential telephone numbers, we can approximate those figures for the 1999 NSAF. Assuming that 4 percent of the roughly 41 million 1999 NSAF cellular-owning telephone households erroneously reported a cellular telephone as an additional residential telephone number, we feel that 1½ to 2 points of the overall 6½ point drop were respondents who included cellular telephones in their count of additional household telephone numbers (4 percent of 41 million). The remaining 4½ to 5 point decline in multiple telephone households from the 1999 to the 2002 NSAF is plausibly the result of both changing telephone technology and reduced demand for more than one household landline.

Table 2. U.S. Households with Additional Residential Telephone Numbers (millions)

Year	Households w/telephone service	Additional residential lines	% additional lines for households w/telephones
1988	85.4	2.3	2.7%
1989	87.4	2.6	3.0
1990	88.4	3.9	4.4
1991	89.4	6.5	7.3
1992	91.0	8.3	9.1
1993	93.0	8.8	9.4
1994	93.7	11.4	12.2
1995	94.2	13.9	14.7
1996	95.1	16.0	16.8
1997	96.5	18.2	18.9
1998	98.0	19.1	19.5
1999	99.1	23.6	23.8
2000	100.2	26.2	26.2
2001	102.2	25.1	24.6
2002 ^a	104.0	18.7	18.0

Note: This table is adapted from table 7.4, “Additional Residential Lines for Households with Telephone Service (End-of-Year Data in Millions),” from the FCC’s August 2003 release of *Trends in Telephone Service*.

a. The 2002 estimate of households with additional telephone lines is an unpublished and preliminary estimated obtained from the FCC.

Other recent RDD surveys also show a decline in households with multiple landlines. Despite some measurement differences, the proportional decline of multiple telephone line households for the period from 1999 to 2002 is relatively similar for the NSAF and BRFSS studies. NSAF data show a 36 percent decline from 1999 to 2002. However, by making the 1½ to 2 point adjustment to the 1999 NSAF figure, we find that the decline drops to 30 percent compared with the BRFSS data, which show a 29 percent decline for the same period. The decline in household demand for supplemental landline numbers is the result of consumer response to advancements in the telecommunications industry. As the percentage of multiple landline households has fallen in recent years the popularity of cellular telephones and digital subscriber lines (DSL) has surged (CTIA 2003). The purchase of cellular telephones may be replacing the purchase of additional landlines in some households. Moreover, DSL Internet connections render the need for supplemental landline Internet numbers unnecessary since they allow for simultaneous voice and Internet communication on a single landline. Comparing the 1999 NHES adult special survey

with the 2002 NSAF, we found that 15 percent more NSAF respondents than NHES respondents say they would answer a telephone number used for computer or fax machine connections (35 versus 50 percent). This change over time can be attributed to increased availability of landline numbers that provide Internet, fax and voice communication simultaneously

5. Summary

As a result of increasing telecommunication options, it has become more difficult to determine a households' chance of selection from a RDD sampling frame. A simple one or two question approach no longer yields the necessary information. Use of an expanded question series to navigate the respondent through the inclusion of cellular telephones lines and landlines used exclusively for purposes other than voice communication is necessary in today's confusing and rapidly changing telecommunications culture. The series of questions asked on the 2002 NSAF survey seems to address these issues, but for how long? In just three years we have seen a 15 percentage point increase in the number of people reporting that they would answer and talk on a telephone number that is also used for computer or fax connections. This demonstrates how rapidly people are changing their telecommunication habits. As the telecommunication world changes it is important to readdress the series of household telephone number questions, keeping in mind that the primary objective is to determine household probability of selection.

Major RDD surveys administered over the past five years have obtained varying estimates of the percentage of multiple telephone number households. Ideally, all telephone surveys would benefit from the development of a set of questions that could become the industry standard. However, this will be a difficult objective to accomplish given the fast paced nature of telecommunication technologies.

References

- Cellular Communications and Internet Association. 2003. "CTIA's Semi-Annual Wireless Industry Survey." http://www.wow-com.com/pdf/CTIA_Survey_Yearend_2002.pdf.
- Federal Communications Commission. 2003. "Trends in Telephone Service: August 2003." http://www.fcc.gov/Bureaus/Common_Carrier/Reports/FCC-State_Link/IAD/trend803.pdf.
- Roth, Shelly B., Jill Montaquila, and Michael J. Brick. 2001. "Effects of Telephone Technologies and Call Screening Services on Sampling, Weighting and Cooperation in a Random Digit Dial (RDD) Survey." *Proceedings of the Section on Survey Research Methods*. Alexandria, VA: American Statistical Association.
- Warren, Patricia, and Pat Cunningham. 2003. *Telephone Survey Methods for the 2002 NSAF*. Methodology Report No. 9.
- Waksberg, Joseph. 1978. "Sampling Methods for Random-Digit Dialing." *Journal of the American Statistical Association* 73(361): 40-46.

An Experiment in Call Scheduling

Patricia Cunningham, David Martin, and J. Michael Brick

1. Introduction

The National Survey of America's Families (NSAF), conducted by the Urban Institute, is part of a multiyear study to assess the new federalism by tracking ongoing social policy reforms and relating policy changes to the status and well-being of children and adults. The major objective of the study is to assess the effects of the devolution of responsibility for major social programs, such as Aid to Families with Dependent Children, from the federal to the state level. The NSAF collects information on the economic, health, and social dimensions of the well-being of children, nonelderly adults, and their families in 13 states and in the balance of the nation. The 13 states, which account for a little more than 50 percent of the country's population, were selected to vary in terms of their size and geographic location, the dominant political party, and key baseline indicators of well-being and fiscal capacity. A sample of the balance of the nation was included so that national estimates could also be produced. Low-income families were oversampled because the policy changes of interest are expected to affect them most. The initial round of the NSAF took place in 1997, with follow-up rounds in 1999 and 2002. The Urban Institute is funded by a consortium of foundations, led by the Annie E. Casey Foundation, to carry out the *Assessing the New Federalism* project. Westat is responsible for NSAF data collection and related activities.

1.1 Summary of Sample Design and Methodology

The survey is a dual frame design with a random digit dialing (RDD) sample to cover the approximately 95 percent of the U.S. population with telephones and an area probability sample to represent households without telephones. A major focus of the survey is to provide reliable estimates for persons and families below 200 percent of the poverty threshold. The area probability sample was included because of concerns about the potential bias in the estimates due to excluding low-income households without telephones.

The RDD portion of the survey used a list-assisted method for sample selection and computer-assisted telephone interviewing (CATI) for screening and interviewing. This component involved screening nearly 200,000 households and conducting detailed 25- to 45-minute interviews with approximately 40,000 to 50,000 persons under the age of 65. The area sample for Round 1 in 1997 and Round 2 in 1999 required listing nearly 40,000 addresses in 138 primary sampling units (PSUs) in order to conduct fewer than 2,000 interviews. For Round 3 in 2002, the state-level area samples were dropped and only a national area sample was conducted. Thus the size of the area sample was reduced by half for Round 3. Cellular telephones were used to connect sample persons in nontelephone households with the telephone center where telephone interviewers administered the questionnaires, reducing any mode effects.

1.2 Experiments in Round 3

During Round 3, a series of experiments were conducted to test incentive levels and timing, income screening, and the automatic call-scheduling algorithm. All the experiments were done in

what we called the predictor sample, a random sample of 60,000 telephone numbers that were fielded before all other sampled numbers. The incentive experiment had six conditions that tested whether a \$2 prepayment at the screener makes a promise of \$10 at the extended more effective than \$5 at refusal conversion at the screener. It also tested whether the promise of \$10 leads to a higher extended response rate than a prepayment of \$5 and a promise of \$20 at refusal conversion. The outcome of this experiment is discussed in Canter et al. (2003).

In the income experiment, alternative methods of screening for income were compared. The purpose was to assess the most accurate means of asking income for oversampling low-income households without being obtrusive and losing the respondent during the screening interview. Accuracy of reporting was crucial for maintaining effective sample sizes. In one option, respondents were asked two or three brief questions to disaggregate their income by thinking about how many people work and about other sources of income. In the second option, respondents were asked if their income was above or below an amount that was well below the low-income cutoff for the study and asked, if necessary, a follow-up question to work up to 200 percent of the poverty level. The final experiment manipulated the automatic call scheduling algorithm. It is the subject of this paper.

2. Background for Experimenting with Call Scheduling

The experiment in call scheduling was focused on the procedures used to contact and determine whether the telephone number was residential. The contact procedure is clearly only one of the scheduling procedures used, but it has important implications for the cost of data collection. Other procedures, such as dealing with numbers that were contacted but not completed, were not tested in this experiment. It is also worth noting that the optimal contact procedure may not be the one that yields the greatest participation in the survey. See Brick et al. (1996) for an example.

In the first round of NSAF in 1997, a call scheduling protocol was used that required the telephone numbers be attempted up to seven times over specific weekday, evening, and weekend periods to establish a first contact. When someone answered the telephone, the residential status for the telephone number was determined. At that point, the noncontact protocol was completed and a different protocol was used, if needed, to obtain the cooperation in the interview. If the telephone number was still not contacted after seven call attempts, it was held for at least a couple of weeks and then rereleased for the same seven-call sequence. Many numbers that were still noncontacts after 14 calls were rereleased for seven more call attempts.

Exploratory analysis of the 1997 NSAF calling records found that the percentage of telephone numbers that result in an initial contact (i.e., it was possible to resolve the residency status as residential or nonresidential) decreased with the number of attempts. An interesting result was that the eighth and the 15th (for those noncontact cases that had more than 14 attempts) call attempts had higher than expected percentage contacts, interrupting the essentially monotonic pattern otherwise exhibited. Since the 1997 scheduling protocol was to make seven attempts and then hold the numbers for additional releases, we hypothesized that the delay between the seventh and eighth and the 14th and 15th attempts was the cause of the higher-than-expected contact rates.

In Round 2 we acted on these findings by revising the calling pattern so that there were delays between the seventh and eighth calls and the eighth and ninth calls. Note that if the hold period was one week, the way the hold was implemented meant that actual time between the calls was at least one week. The total number of attempts was reduced from the minimum of 14 used in Round 1 to nine for Round 2. A subsample of telephone numbers was fielded for additional calls to support estimating the residency rate using the survival method.

In Round 3 we revised the scheduling pattern for noncontact numbers to take fuller advantage of these results. We hypothesized that we might achieve the same increase in contacts with even fewer call attempts. The Round 3 approach involved an initial four attempts (over weekdays, weekends, and evenings), a one-week hold period, an additional three attempts (completing the requirements for spread of the calls as used in Rounds 1 and 2), a one-week hold, and then a release for two additional call attempts. The goal was to more quickly contact households and have fewer numbers in noncontact status. This approach would be used for most of the sample, with a subsample of numbers assigned a greater number of attempts for the survival estimation.

To evaluate the effectiveness of this new approach, we conducted an experiment using a more traditional approach of seven initial calls, a one-week hold, and two additional call attempts. The experiment is described in more detail in the next section.

3. Scheduler Experiment in Round 3

The hypothesis for this experiment stated that if we reduced the number of calls from seven to four before we held the case for a week, we would reduce the number of call attempts required to determine residency. Essentially we expected to find the percentage contacted would increase on the fifth call attempt due to the introduction of the hold period. There were two conditions, named 4.3.2 and 7.2, with the dot representing the hold period.

- *The 4.3.2 Condition.* In this condition, the initial four calls to establish contact were automatically scheduled so that one fell on a weekend (either Saturday or Sunday), one during a weekday (9 AM to 6 PM), and two on weekday evenings (6:00 to 7:30 and 7:30 to 9:00) within the respondent's time zone. The calls could be made in any order. If there was no contact, it was held for one week before it was rereleased for an additional series of three calls, one on the weekend, one during a weekday, and one on a weekday evening (6 PM to 9 PM). If, for example, during the first four calls, an early call (9 AM–2 PM) was made, then we made the other day call (2 PM–6 PM) as part of the remaining three calls. If we called Saturday during the first four calls, we called Sunday during the next three calls. Again the calls could be made in any order. If there was no contact after seven calls, it was held for another week at which point it was rereleased for an additional two calls, one in the evening (6 PM to 9 PM) and one on a weekend.
- *The 7.2 Condition.* In this condition, the initial seven calls to establish contact were automatically scheduled so that there were two on the weekend, two during weekdays (9 AM to 6 PM), and three during weekday evenings (6:00 to 7:30, 7:30 to 9:00, and 6:00 to 9:00) within the respondent's time zone. These calls could be placed in any order. If there was no contact after seven calls, it was held for another week at which point it was rereleased for an additional two calls, one in the evening (6 PM to 9 PM) and one on a weekend.

Notice that after seven attempts, the number and placement of calls is exactly the same in both conditions, e.g., two during weekdays, two on weekends, and three on weekday evenings. The only difference is the one-week hold period between the fourth and fifth calls in the 4.3.2 condition.

Within the predictor sample of 60,000 telephone numbers, we predesignated a subsample of 5,000 telephone numbers to implement the more traditional approach with seven initial calls, a one-week hold period, and two additional call attempts (7.2). The remaining 55,000 numbers received the 4.3.2 calling algorithm. Other experiments on income questions and incentives were also carried out using the predictor sample, but a factorial design was used so the sample sizes for the other experiments were balanced for the 4.3.2 and 7.2 conditions. Thus, the other experiments in no way interfered with the scheduler experiment.

4. Findings

Before examining the results of the experiment in terms of the main outcome variable (the number of call attempts to determine residency status), we present some basic results for the two experimental groups. Table 1 shows the sample of 5,000 was randomly assigned to the 7.2 condition and the remaining 55,000 numbers were assigned to the 4.3.2 condition. The percentages in the table show that the two experimental groups were very similar with respect to the number of telephone numbers that were purged (eliminated from dialing because they matched to a business number or were nonworking when they were autodialed). The percentage of numbers that were completed and the percentage of numbers identified as residential were also very similar for the two experimental groups. The differences in the percents for the two groups shown in the last column are all less than the standard error of the difference, so differences this large could be expected by chance. Since the telephone numbers were randomly assigned to the groups, the data in table 1 essentially verify the random assignment process.

Table 1. Experimental Groups Outcomes

	Number		Percent		Difference
	7.2 group	4.3.2 group	7.2 group	4.3.2 group	
Sample size	5,000	55,000	100.0	100.0	
Purged	1,892	21,149	37.8	38.5	-0.6
Dialed	3,108	33,851	62.2	61.5	0.6
Contacted	2,643	28,740	52.9	52.3	0.6
Completed	1,292	14,300	25.8	26.0	-0.2
All residential	1,874	20,545	37.5	37.4	0.1

Note: The standard error of the difference of percents is approximately 0.7, so no differences are significant.

Now we examine the distribution of the percent of numbers with resolved residential status by the number of call attempts to determine if the expected relationship holds for the two experimental groups. For this and all further analysis, we only include those telephone numbers resolved by the ninth call attempt. Figure 1 graphs the cumulative percentage of all sampled telephone numbers that have a resolved residency status for each group. The graph shows that after the first call, the percent resolved is about the same for the two groups, but the difference between the 4.3.2 and 7.2 groups is statistically greater than zero at both the second and third call

attempts. The difference decreases as the number of call attempts increases. There is no indication of the desired increase in the percent resolved at the fifth attempt for the 4.3.2 group.

Before exploring the unexpected differences in the percent resolved shown in the second and third call attempts, we more directly examine the anticipated increase in the conditional probability of contact. Figure 2 shows the conditional probability of determining the residential status of a telephone number for the two groups by the number of call attempts. For example, the conditional probability of resolving a telephone number on the second attempt is the number of telephone numbers that were resolved on the second attempt divided by the number that required at least two calls.

The typical pattern is that the conditional probabilities of resolving the numbers decrease with the number of attempts. Figure 2 shows this general pattern but there are some important features. The conditional probabilities for the 4.3.2 group are substantially greater than the same probabilities of the 7.2 group at the second and third call, resulting in the differences noted in figure 1. The conditional probabilities for both groups are also not monotonic, and differences from the decreasing rate are clearly seen in the third and eighth call attempts.

Focusing on just the 4.3.2 group, we do not see a spike in the conditional probability of resolving a telephone number at the fifth call, as had been hypothesized. In contrast, consider the eighth attempt that was also delayed for a week, does have a spike that is especially apparent for the 4.3.2 group. If the numbers were not held for at least one week between call attempts seven and eight, we would have expected a constant or decreasing conditional probability of resolving at this point. Figure 2 clearly shows that the anticipated relationship in terms of the changes in conditional probabilities for the 4.3.2 group were not obtained, nevertheless the differences in the resolution rates in the early calls are very interesting and these are considered next.

When we examined figure 1 and saw the differences in the resolution rates at the second and third attempts, we began to investigate the possible source of these differences. One way is to classify the telephone number by the residency status eventually assigned for the number. While we classified the numbers by a number of categories, we only present the one that was most interesting here. Figure 3 is constructed exactly like figure 1, except it is restricted to telephone numbers that were determined residential. Figure 3 shows that residential numbers contributed to the differences in the resolution of the telephone numbers in the second and third call attempts. The resolution rates for the two groups do not converge until the fifth call or later. In fact, most of the differences in figure 1 are due to the contribution of residential numbers, as displayed in figure 3.

Figure 1. Distribution of All Numbers Resolved, by Call Attempt and Experimental Group

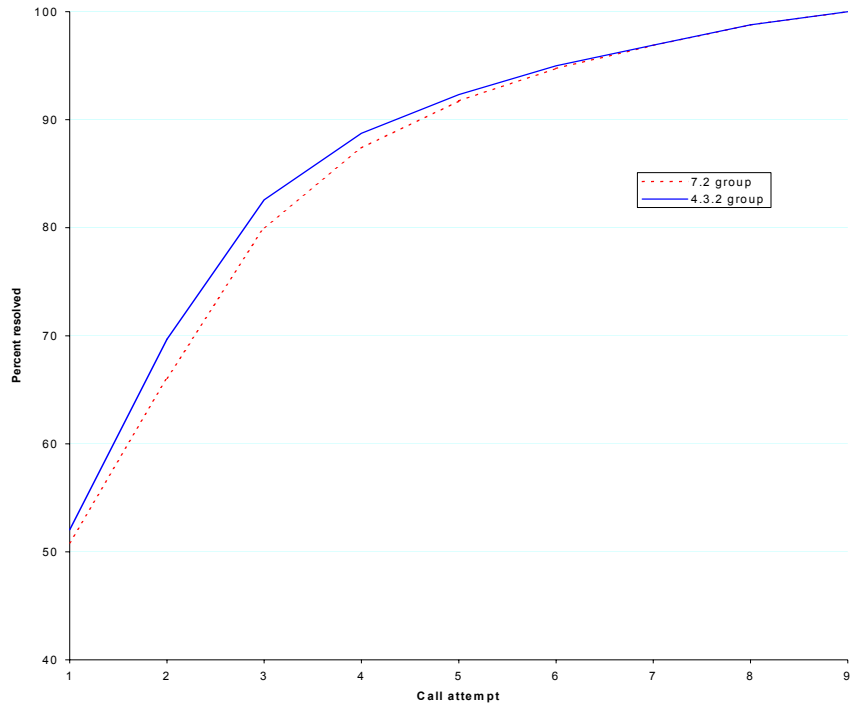
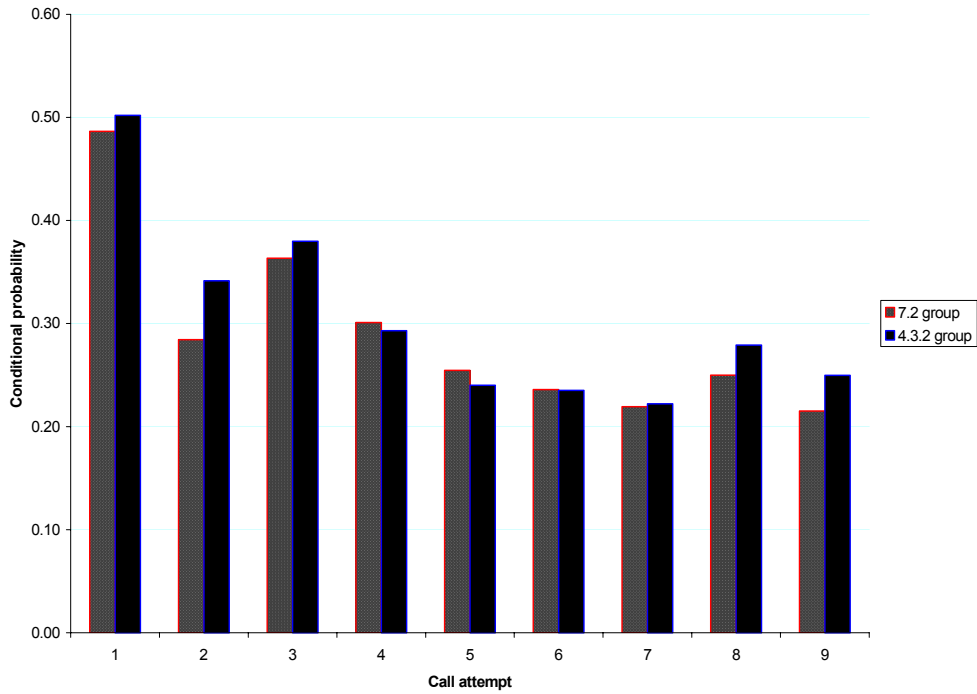


Figure 2. Conditional Probability of Resolving Residency Status on a Call Attempt, by Group



Another way of categorizing the telephone numbers is by whether an address could be found for the telephone number that could then be used to send an advance letter to the household. Those numbers associated with an address are called *mailable*. This classification was considered because *mailable* numbers have much higher residential rates than *nonmailable* numbers. When we examined the distributions by *mailable* and *nonmailable* status, we found that *mailable* status did not account for the differences in the experimental groups. The sample size for the 7.2 *nonmailable* group was too small to do a detailed analysis. However, looking at all 60,000 telephone numbers by *mailable* status is interesting in its own right. Figure 4 gives the cumulative percentage of all sampled telephone numbers that have a resolved residency status for *mailable* and *nonmailable* telephone numbers. The graph shows that *mailable* numbers are contacted and resolved with fewer attempts than *nonmailable* numbers. *Mailable* numbers are much more likely to be residential than *nonmailable* numbers, so we might expect a slightly lower number of call attempts for the *mailable* numbers (overall, residential numbers required on average 2.3 call attempts to resolve, while nonresidential numbers required 2.4). The mean number of attempts for *mailable* numbers was 2.2 and for *nonmailable* numbers was 2.5. This finding adds some new evidence about the relative efficiency of sampling by *mailable* status (Brick et al. 2002).

While these results are interesting, they do not explain the difference in the rates of resolving residency status for the two experimental groups at the second and third call attempts. To address this concern, we decided to evaluate the calling patterns for the two groups. We understood that even though the distribution of call attempts by time period (day/evening/weekend) for the 4.3.2 and 7.2 had to be equivalent after seven attempts, they were not forced to follow the same path to this point. In fact, the 4.3.2 group calling pattern was likely to be different from the pattern for the 7.2 group, because numbers assigned to the 4.3.2 group were required to have one day, one weekend, and two evening attempts in the first four calls. This restriction was not placed on the 7.2 group. Clearly, this could have an important effect on the rate of resolving the cases.

Table 2 gives the percent of telephone numbers attempted in each experimental group by calling pattern for the first four call attempts. The percentages do not always add to 100 because any pattern that was less than 0.5 percent for both the 4.3.2 and the 7.2 group was excluded from the table. The last column gives the percentage point difference between the estimates for the 4.3.2 group and the 7.2 group. For the first call attempt, the calling patterns are essentially the same for the two groups as should be expected since no restrictions were placed on the first attempts.

On the second attempt, the differences are substantial. The 7.2 group is much more likely to have two attempts on the weekend or during the day, while this pattern was generally not permitted for the 4.3.2 group. The scheduling algorithm did allow some exceptions to the calling rules if it was necessary for workflow reasons. As a result, the percentages in the undesirable patterns are small but greater than zero. For those numbers that required a third attempt, the call distribution for the two experimental groups is still very different. Nearly all the numbers in the 4.3.2 group have one day attempt and one or two evening attempts. On the other hand, over 83 percent of the numbers in the 7.2 group do not have an evening attempt. Even after the fourth attempt, about 25 percent of the 7.2 group have no evening call, while virtually all the numbers in the 4.3.2 group have two evening calls at the same point. Since the rate of contact for residential numbers is

much greater in the evening than during other time periods, the call pattern appears to largely account for the large differences in rates for the two groups overall and for residential numbers.

The findings in Table 2 show that the change in the scheduling algorithm did affect the ability to contact households and determine residential status, but the consequences were not exactly what were expected. The 4.3.2 algorithm required a distribution of call attempts by time period that increased the rate for contacting households in fewer attempts. As the number of call attempts increased, the probability of having a less-desirable pattern for the 7.2 group decreased, and by the seventh attempt the two groups had equivalent distributions by time of attempt. This convergence corresponds to the convergence of the cumulative percentage of cases resolved by call attempt for the two experimental groups.

Figure 3. Distribution of Residential Numbers Resolved, by Call Attempt and Experimental Group

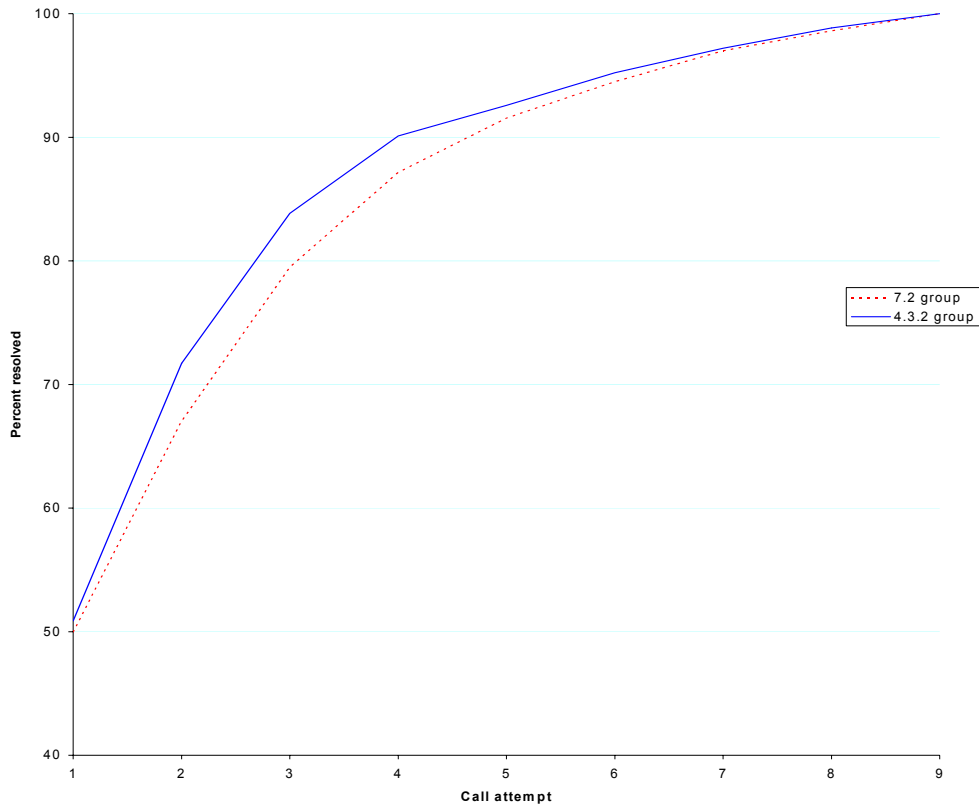


Figure 4. Distribution of Percentage of all Numbers Resolved, by Call Attempt and Mailable Status

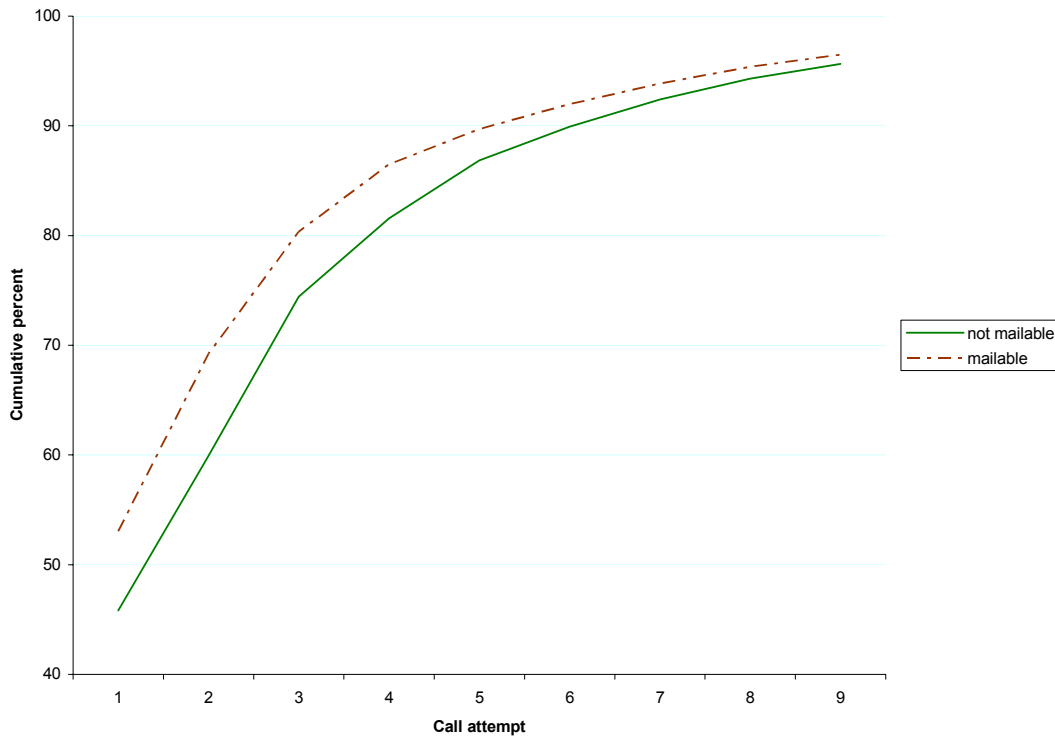


Table 2. Time of Attempts, by Experimental Groups

	Time of call			Percent of attempted		Difference
	Day	Evening	Weekend	7.2 group	4.3.2 group	
1st attempt	0	0	1	37.1	37.2	-0.2
	0	1	0	15.9	15.1	0.8
2nd attempt	1	0	0	47.1	47.7	-0.6
	0	0	2	19.1	1.7	17.4
	1	0	1	26.4	48.2	-21.8
3rd attempt	2	0	0	14.2	45.3	-31.1
	2	0	0	37.6	2.5	35.1
	1	0	2	24.8	1.6	23.2
	1	1	1	4.6	70.2	-65.6
	1	2	0	0.4	24.4	-24.0
4th attempt	2	0	1	23.1	1.9	21.2
	2	1	0	40.2	1.7	38.5
	1	1	2	11.8	2.1	9.7
	1	2	1	0.4	92.5	-92.2
	2	0	2	26.5	0.4	26.1
	2	1	1	50.3	2.9	47.4
	2	2	0	6.1	0.9	5.2

Notes: Excludes patterns with less than 5 percent of all attempts in any group. All differences are statistically significant except those for the 1st call attempts.

5. Discussion

The scheduler experiment compared two different calling algorithms to evaluate a hypothesis that one of the approaches would increase the probability of reaching the household and determining if the number was residential or not. The 4.3.2 procedure required holding telephone numbers at least one week after the fourth call, expecting that this would increase the conditional probability of resolving the status on the fifth call as compared to the 7.2 procedure that did not have this hold period. The findings showed that using the 4.3.2 procedure did not increase the conditional probability of resolving a number at the fifth call. However, the 4.3.2 procedure was superior to the 7.2 procedure in an unanticipated way.

The 4.3.2 procedure required dialing with calling patterns by time period (combinations of day/evening/weekend calls) that had a higher likelihood of contacting households. In particular, the first four calls in the 4.3.2 procedure had to have one day, one weekend and two evening attempts. The 7.2 procedure did not have this requirement and was much more likely to have few evening calls in the first few dialing attempts. Since the distribution of the first seven calls by time period for the 4.3.2 procedure and the 7.2 were required to be the same, the difference in the percentage of numbers resolved decreased as the number of call attempts increased.

The experiment clearly shows that scheduler algorithm that requires a variety of time periods in early calls reduces the total number of call attempts needed to contact households and to eliminate nonresidential numbers. In a large survey that extends over a relatively long calendar period like the NSAF, this type of scheduler is feasible and recommended. Since the 4.3.2 procedure was used for all the nonexperimental telephone numbers in NSAF in 2002, the benefits were obtained in the survey.

On the other hand, the 4.3.2 scheduler algorithm and others that place greater restrictions on when calls to a number can be made do have implications for survey operations. As more restrictions are placed on when calls can be made, the flow of cases to the interviewers may be affected. In the extreme, a scheduler that places too many restrictions may cause interviewers to have no work to call at certain time periods. Clearly, this would be much less efficient than allowing calls to numbers at less than optimal periods. This type of situation is most likely to happen in surveys that have a short field period or have limited interviewing resources. In such cases, more sophisticated algorithms that prioritize numbers to be dialed so those most likely to benefit from a call attempt at the current period are scheduled first are needed. In fact, the scheduler used for NSAF actually had many of these features.

References

Brick, J.M., B. Allen, P. Cunningham, and D. Maklan. 1996. "Outcomes of a Calling Protocol in a Telephone Survey." *Proceedings of the Survey Research Methods Section of the American Statistical Association* (142–49). Alexandria, VA: American Statistical Association.

Brick, J.M., D. Judkins, J. Montaquila, and D. Morganstein. 2002. "Two-Phase List-Assisted RDD Sampling." *Journal of Official Statistics* 18:203–16.

Cantor, D., P. Cunningham, T. Triplett, and R. Steinbach. 2003. "Comparing Incentives at Initial and Refusal Conversion Stages on a Screening Interview for a Random Digit Dial Survey." Presentation at AAPOR, Nashville, TN.

Increased Efforts in RDD Surveys

Michael Brick, David Martin, and Patricia Warren, Westat; J. Wivagg, DataSource

1. Introduction

The Urban Institute conducts the National Survey of America's Families (NSAF) to assess the new federalism by surveying children and adults in 13 states and in the balance of the nation. The 13 states account for about 50 percent of the country's population. The initial round of the NSAF took place in 1997, with follow-up rounds in 1999 and 2002. The major objective of the study is assessing the effects of the devolution of responsibility for major social programs. A consortium of foundations, led by the Annie E. Casey Foundation, funds the NSAF. Westat is responsible for the sample design, data collection, and weighting. The survey is a dual frame design with a random digit dialing (RDD) sample and an area probability sample. In this paper, we consider only the RDD sample. See Cunningham, Martin, and Brick (2003) and the Urban Institute's web site (www.urban.org) for more details on the survey and its methodology.

Recent literature on random digit dial (RDD) telephone surveys suggests that increased levels of effort are required to contact households due to changes in technology (see, for example, Curtin et al. 2000). Answering machines, caller-ID devices, TeleZappers, and privacy managers are examples of the types of technologies that are being used at an increasing rate to shield households from unwanted calls. In addition, it is speculated that more telephone numbers are devoted to purposes such as fax machines and computers and may never be answered no matter how many times the numbers are dialed. This paper uses Round 1 (1997) and Round 3 (2002) NSAF data to examine some of these issues. The next section gives some basic background information needed to help understand the data collection efforts in the two rounds.

2. Background

As noted earlier, the three rounds of data collection for NSAF were done in 1997 (Round 1), 1999 (Round 2), and 2002 (Round 3). In all three efforts, the content and the method of interviewing for the RDD sample were very similar.

The Round 1 and Round 3 RDD sample designs were also very similar. For these two rounds, independent, list-assisted RDD samples were selected from the study areas and from the balance of the nation. The sample design for Round 2, on the other hand, differed from this design. The Round 2 sample included a substantial subsample from Round 1 and the telephone numbers were differentially sampled depending on the Round 1 outcome. In addition, a supplemental sample of new telephone numbers was included in Round 2 to provide complete coverage and reach the desired sample size for that round. Because the Round 2 design is so different from the other two rounds, we only consider Round 1 and Round 3 in this paper.

While the basic sample designs for Round 1 and Round 3 were similar, there were some important differences that affect any comparisons of levels of effort. One change was in the composition of the study areas and the sample distribution. In Round 1, Milwaukee and the balance of Wisconsin were separate study areas with their own RDD samples, but in Round 3 Milwaukee was merged with the balance of Wisconsin to form a study area with one RDD

sample. Another difference was that the sample allocated to the balance of the nation was larger in Round 3 than in Round 1.

Other important differences in procedures were implemented over the three years between Round 1 and Round 3. Many of these are discussed below, but two that have very important implications for our analysis are mentioned here. The first procedure is the use of incentives. In Round 1, an advance letter was sent to each telephone number linked to a mailing address, but monetary incentives were reserved for converting households and sample persons who refused. In Round 3, a \$2 prepaid incentive was included for telephone numbers linked to a mailing address. If a household was sampled for the extended interview, it was promised \$10. It is also worth noting that the ability to find an address for a telephone number has increased over time. In Round 1, only 38 percent of the sampled telephone numbers had mailable addresses, while in Round 3 the share increased to 72 percent. Of course, the mailable addresses may not reach the household for the sampled telephone number due to errors of various sorts. Previous research suggests that 15 to 20 percent of the mailable addresses may not be accurate.

A second procedure that affects comparisons is subsampling households that refused to participate in the screening interview. In Round 3, households that refused to respond to the screening items were subsampled and only those subsampled were subject to refusal conversion efforts. About 4 percent of all telephone numbers were screeners that refused and were not subsampled for refusal conversion. This accounts for about 11 percent of all the residential numbers in Round 3. The subsampling refusal procedure was not used in Round 1. We comment on both these procedures in more detail as we examine the outcomes.

The tabulations and comparisons of the outcomes from Round 1 and Round 3 presented below are all based on the actual counts and have not been weighted. The advantage of using the raw counts is that it corresponds to the actual levels of effort expended in the operations at the time, which is the focus of the paper. The main disadvantage is that changes in the design such as those described above are more likely to contribute to differences between the rounds. For example, the change in the composition of the sample (Wisconsin and balance of the nation samples) would not be as relevant in weighted analyses.

We begin by giving some basic comparisons of the RDD samples selected for the two rounds. Table 1 gives the number of telephone numbers sampled by residential status for each round. As shown in the table, the number of telephone numbers sampled was much greater in Round 3, partially because the residency rate of the sampled numbers in list-assisted RDD samples decreased rather dramatically in the five-year period. For the NSAF samples, the percent of telephone numbers sampled that were residential (excluding the telephone numbers with a residency rate that could not be determined) decreased by 9 percentage points, a decrease of 91 percent. Another notable difference was the increase in the percent of sampled telephone numbers that had an unknown residential status. This group increased less than 3 percentage points, but this change is large compared to the percent of numbers in the category. The other factor that must be taken into consideration is that the number of telephone numbers in the list-assisted frame also changed drastically. The relative difference in the frame increased by 30 percent from Round 1 to Round 3. This increase makes it difficult to interpret the percentage change in the unknown residential numbers.

Table 1. Disposition of Sampled Numbers, by Residency Status

Disposition	Round 1	Round 3	Difference	Relative difference
Number sampled	483,260	556,651	73,391	15.2%
Residential	46.0%	36.9%	-9.1%	-19.9%
Nonworking	32.9	40.2	7.3	22.2
Nonresidential	15.0	14.0	-0.9	-6.1
Unknown residency status	6.1	8.8	2.8	45.1

Note: Numbers may not add due to rounding.

The much lower residency rate in Round 3 compared with Round 1 might suggest that the level of effort needed to reach households must have increased substantially in these five years. However, survey researchers also used new technologies to deal with the increased proportion of sampled numbers that are not residential. In both rounds, the sampling vendor we used, Marketing Systems Group (MSG), purged the sample of telephone numbers of numbers only listed in the Yellow Pages and then autodialed other numbers to eliminate nonworking and some business numbers from the sample. The MSG purging technology (called ID in 1997 and ID*plus* in 2002) advanced so that 38 percent of the sampled telephone numbers was purged in Round 3, while only 18 percent was purged in Round 1. Of the sampled telephone numbers that were not purged, the percentage that were residential increased by about 3 percentage points from Round 1 to Round 3. Thus, the lower residential rate in Round 3 shown in table 1 did not increase the percentage of numbers that had to be dialed in the survey because of this technological innovation. In the subsequent sections the purged telephone numbers are not included.

3. Findings

This section analyzes and compares the levels of effort from the Round 1 and Round 3 surveys. The analysis is restricted to the screening interview where the initial contact is made and any sampling of persons within the household is conducted. In both rounds, the content of the screener was basically the same. The first analysis examines the number of call attempts to complete the screener. The second analysis looks at telephone numbers that ring and are never answered in an effort to find out if more telephone numbers are used for technological purposes. The third analysis looks at levels of nonresponse over time and how other procedural changes affect these outcomes.

3.1 Call Attempts

The number of call attempts to finalize a case is a common measure of a survey's level of effort. Table 2 gives the mean number of call attempts for the two rounds of data collection by the residency status of the number. Overall, the number of call attempts increased by 10 percent between rounds, but this increase is somewhat deceptive because of some procedural differences mentioned above. In particular, the Round 3 refusal cases that were not retained due to subsampling are included in the tabulation, although calls were not made to these numbers after the initial refusal. This is an important factor that complicates the analysis. Further, the decrease in the mean number of attempts to numbers with unknown residency status is directly attributable to procedural changes. Cunningham et al. (2003) describe these changes and the rationale for them.

When we examine the mean number of call attempts for the residential numbers, the problem associated with not accounting for the refusal subsampling becomes even more apparent. In table 3, the mean numbers of attempts for all numbers classified as residential are given by the final disposition of the screener. Most categories are obvious, except perhaps Max Calls. These are households that answer the phone at least once, never refuse, but never are available to complete the interview despite repeated call attempts. The table suggests that every disposition, except other nonresponse, has a smaller number of call attempts in Round 3 than in Round 1. Since the other nonresponse group is very small (less than 0.1 percent of all dispositions), it is obvious that understanding the effect of the level of effort requires dealing with the refusal subsampling.

Table 2. Mean Number of Call Attempts, by Residency Status

Final Disposition	Round 1	Round 3	Difference	Relative difference
Total	6.2	6.8	0.6	9.7%
Residential	6.9	7.1	0.2	2.9
Nonworking	3.0	3.1	0.1	3.3
Nonresidential	3.4	4.4	1.0	29.4
Unknown residency status	15.9	12.0	-3.9	-24.5

Note: Numbers may not add because of rounding.

Table 3. Mean Number of Call Attempts for Residential Numbers, by Final Disposition

Disposition type	Round 1	Round 3	Difference	Relative difference
Complete	4.8	4.5	-0.3	-6.3%
Language problem	14.6	13.5	-1.1	-7.5
Max call	37.9	24.3	-13.6	-35.9
Refusal	13.6	10.3	-3.3	-24.3
Other nonresponse	3.5	11.8	8.3	237.1

Note: Numbers may not add because of rounding.

Table 4 divides the residential numbers into those that never had a refusal and those that had one or more refusals to clarify the effect of refusal subsampling. In the top of the table, the number of call attempts for the never-refused numbers are relatively consistent across rounds. The only notable difference is that more attempts were made in Round 1 than in Round 3 before classifying the number as a Max Call. A procedural decision was made in Round 1 to set the number of call attempts before finalizing a case as a Max Call to a very large number. In Round 3, the call attempt limit was reduced. Later we show that the percentage of cases classified as Max Calls more than doubled from Round 1 to Round 3. Since this group of numbers required more attempts, the calling in Round 3 actually increased. Further, even though the level of effort to resolve the Max Calls was greater in Round 1, the literature shows that once the number of attempts exceeds 20, there is little increase in the number of completed interviews.

Table 4. Mean Number of Call Attempts for Residential Numbers, by Refusal Status

Disposition type	Round 1	Round 3	Difference	Relative difference
Never-refused cases				
Complete	3.8	3.8	0.0	0.0%
Language problem	13.7	12.6	-1.1	-8.0
Max call	37.9	24.4	-13.5	-35.6
Ever-refused cases				
Complete	7.4	10.0	2.6	35.1
Language problem	15.6	15.0	-0.6	-3.8
Refusal	13.6	10.3	-3.3	-24.3
Not subsampled refusal	13.6	14.1	0.5	3.7

Note: Numbers may not add because of rounding.

The lower portion of table 4 gives the means for screeners with at least one refusal. The table shows that it took 2.6 more call attempts to obtain a completed screener in Round 3 than in Round 1 (an increase of 35 percent). The row showing the lower mean number of attempts for the refusal cases is again a function of the refusal subsampling. The last row of the table gives the comparable group by excluding those cases that were not retained in the refusal subsampling. It shows that the average number of call attempts was nearly the same for the two rounds for the refusal cases, giving quite a different picture than from the earlier table. Overall, the mean number of call attempts in Round 3 is greater than the number in Round 1.

3.2 Unanswered Numbers

The next analysis deals with telephone numbers classified as unknown residency status numbers in table 1. These are numbers that are dialed numerous times but never answered. These numbers are partitioned into two groups: Never Answered (NA) numbers are those that ring and are never answered across all call attempts. Answering Machine (AM) numbers are those that are never answered by a person but an answering machine is encountered in one or more call attempts. The procedure we followed was to not depend on an interviewer's interpretation of the answering machine message to classify the number as residential or not.

Table 5 gives the percent of all numbers attempted classified as NA and AM in the 1997 and 2002 NSAF. The table shows that percent of unanswered numbers increased substantially and both the NA and AM classifications grew, but NA numbers account for most of the unanswered numbers. The literature suggests that answering machines are not a major problem if enough calls are made to the number (e.g., Oldendick 1993; Oldendick and Link 1994). The Round 1 results are consistent with this finding, but by Round 3 the percent of AM numbers increased, and the percentage in this category became more problematic.

One factor hypothesized to account for part of the large increase in the number of sampled telephone numbers that are not residential (see table 1) is households dedicating lines for fax or computer use. The increase in the percent of unanswered numbers suggests this possibility. However, as noted earlier the frame of all numbers also changed during this period. Although we have no direct evidence, we would expect a higher percentage of call attempts to such dedicated numbers to be either 'ring no answer' or busy outcomes (if the fax or computer was in use). To investigate this, we examined the NA call attempts in the two rounds. Since both rounds had at

least nine call attempts for every NA (in Round 1 all had at least 14 attempts and in Round 3 all had at least nine), we only look at the results of the first nine call attempts for comparability.

Table 5. Percent of Dialed Numbers That Were Never Answered, by Round

Never answered	Round 1	Round 3	Difference	Relative difference
Total	6.1%	8.9%	2.8%	45.9%
NA	5.2	7.3	2.1	40.4
AM	0.9	1.6	0.7	77.8

Note: Numbers may not add because of rounding.

When we reach a busy number in the NSAF we automatically set an appointment in the call scheduler to dial that number again 15 minutes later. Up to four consecutive busy outcomes are allowed with this procedure. Our usual practice is to count these busy dialings as one call attempt (this is the method used in the earlier tables). For this analysis, we count each dialing separately. Thus, in nine call attempts we could have up to 36 busy dialings. Table 6 presents some characteristics of the distribution of the busy dialings for the NAs in the two rounds. The mean number of busy dialings to NAs doubled from Round 1 to Round 3 (1.4 to 2.9). The percentage of telephone numbers in which half or more of the dialings were busy also doubled. Thus, these data are consistent with the hypothesis of increased use of numbers for other purposes. The implications are twofold. First, these numbers require a large number of dialing attempts, do not result in completed interviews, and increase the cost of data collection. Second, some of the numbers with a NA disposition are usually allocated as residential for computing response rates. If many of these are dedicated for computer or fax use but are counted as residential, then we may be underestimating response rates in RDD surveys.

Table 6. Distribution of Busy Dialings for NA Telephone Numbers

At least ...busy dialings	Round 1	Round 3	Difference	Relative difference
1	17.0%	24.6%	7.6%	44.7%
2	10.1	20.5	10.4	103.0
3	8.5	18.8	10.3	121.2
18	3.2	6.8	3.6	112.5
30	1.4	3.3	1.9	135.7
Mean	1.4	2.9	1.5	107.1

Note: Numbers may not add because of rounding.

3.3 Response Issues

Households that refuse account for the vast majority of all screener nonresponse in the NSAF and nearly all RDD surveys. In Round 1, 89 percent of nonresponse was due to refusals, with Max Calls and language problems accounting for 6.5 percent and 4.0 percent, respectively. Similarly in Round 3, refusals were 86 percent of all nonresponse (including the cases that were subsampled), Max Calls were 10.0 percent, and language problems were 3.9 percent. As noted earlier, the Max Calls require the greatest number of call attempts and the increase of 177 percent in these numbers has cost and response rate implications. However, because the percent of nonresponse due to refusals is so large, this category of nonresponse is the main problem and is considered now.

There are two ways to examine the effect of refusals over time using NSAF data without having to deal with refusal subsampling in Round 3. The first is to examine the percentage of residential numbers that ever refused (initial refusal rate). The second is to examine the percentage of refusals classified as hostile (most of these are classified as hostile on the first contact and thus are not affected by subsampling). Table 7 gives the percent of residential numbers that ever refused and the percent of refusals interviewers classified as hostile in the two rounds. Perhaps the most striking result in the table is that the percentages do not vary much across rounds.

Table 7. Percent of Residential Numbers That Refused and Percent of All Refusals That Were Hostile

 Screener outcomes	 Round 1	 Round 3	 Difference	 Relative difference
Initial refusal	45.4%	47.3%	1.9%	4.2%
Hostile refusals (of all refusals)	1.0	0.8	-0.2	-20.0

Note: Numbers may not add because of rounding.

The refusal rates in table 7 may appear unexpected, especially since the screener response rate was 77 percent for Round 1 and 65 percent for Round 3 using a weighted version of AAPOR (2000) definition RR3. This relatively large decrease in the response rate is generally not consistent with a constant initial refusal rate. The difference relates to how incentives were used in the two rounds. In Round 3, the incentive was front-loaded in the advance letter, because evidence suggested doing this might lower the initial refusal rate. In Round 1, incentives were used to convert those households that refused, thus boosting refusal conversion rates. As a result, the refusal conversion rate in Round 1 was 49 percent and in Round 3 was only 38 percent (for those subsampled for conversion). If the prepaid incentives had not been used in Round 3, it is likely (see the experimental data on incentives in Round 3 in Cantor et al. 2003) that the initial refusal rate for Round 3 would have been 6 percentage points higher. This result is more consistent with the overall screener response rate and the common perception of the research community that refusals are becoming an even greater problem in RDD surveys over time.

4. Summary

The findings above show that while much has changed in only five years, it is difficult to evaluate the effect that would have been obtained if the same procedures were used uniformly over time in the same survey. The NSAF is a good example in the sense that the interview itself had only small changes, but the procedures used were constantly revised to attempt to keep up with changes in the RDD survey environment.

The data show that despite the increasing percentage of the frame that is not residential, revisions in methods used by researchers to purge the frame of nonresidential numbers more than keep pace with this change. We did find the mean number of call attempts increased from Round 1 to Round 3, despite changes in calling protocols that limited unproductive calls to numbers that virtually never resulted in a completed interview. The comparisons for residential numbers were somewhat difficult to analyze due to the introduction of refusal subsampling in Round 3. The main points that emerged were the mean number of attempts for households that did not refuse were constant over time, the mean number of attempts went up substantially for those

households that completed after refusing at least once, the mean number of attempts did not change much for the households that ended up in the refusal category, and the overall mean number of attempts increased also because the number of Max Call cases increased.

We also looked at the NA cases to try to assess if there was any support to the notion that more telephone numbers are being devoted to fax and computer uses. By looking at the busy dialings to these numbers, we found nearly twice as many appeared to be dedicated to uses other than regular incoming and outgoing calls. We concluded that this change requires more dialings, but does not increase the completion rate. We also suspect it may be artificially depressing response rates in RDD surveys.

The third topic we examined was initial refusal rates and found that the procedural changes made this difficult to evaluate. While NSAF response rates declined over the five years, the initial refusal rates remained about the same because incentives were used differently. We believe that the Round 3 procedure enabled us to achieve higher response rates than would have been possible for the same cost using the Round 1 methods.

In general, the statement that increased efforts are necessary to achieve good results in RDD surveys is not very relevant. The environment for RDD surveys changes rapidly, and methods to contact and obtain completed interviews must be continually revised to address those changes. The NSAF is an example of attempts to do exactly that and provide valuable data to analysts.

References

- AAPOR. 2000. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. Ann Arbor, MI: AAPOR.
- Cantor, D., P. Cunningham, T. Triplett, and R. Steinbach. 2003. "Comparing Incentives at Initial and Refusal Conversion Stages on a Screening Interview for a Random Digit Dial Survey." Presentation at AAPOR in Nashville, TN.
- Cunningham, P., D. Martin, and J.M. Brick. 2003. "Scheduler Experiment." Presentation at AAPOR in Nashville, TN.
- Curtin, R., S. Presser, and E. Singer. 2000. "The Effects of Response Rates Changes on the Index of Consumer Sentiment." *Public Opinion Quarterly* 64:413–28.
- Oldendick, R. W. 1993. "The Effect of Answering Machines on the Representativeness of Samples in Telephone Surveys." *Journal of Official Statistics* 9:663–72.
- Oldendick, R.W., and M. Link. 1994. "The Answering Machine Generation. Who Are They and What Problem Do They Pose for Survey Research?" *Public Opinion Quarterly* 58:254–73.

Comparing Incentives at Initial and Refusal Conversion Stages on a Screening Interview for a Random Digit Dial Survey

*David Cantor and Patricia Cunningham, Westat;
Timothy Triplett and Rebecca Steinbach, The Urban Institute*

1. Problem

This paper reports the results of an experiment that tests differences in the timing of the delivery of incentives. More specifically, it compares sending prepaid incentives at the initial contact call to sending advance incentives to households that initially refuse to cooperate. In a paper presented at last year's AAPOR meetings, we reported results from an experiment comparing these two procedures for a relatively small, geographically restricted, sample. In this paper, we provide the comparisons for a national sample (Cantor, et al., 2002).

Three research questions are addressed in this paper:

1. What is the effect of a prepaid incentive of \$2 at the initial contact compared with a \$5 incentive sent at refusal conversion?
2. Do the incentive treatments differentially affect population subgroups?
3. Do the incentive treatments affect data quality?

The amount of the incentive provided in these two alternative procedures (\$2 at initial versus \$5 at refusal conversion) is not the same. If one wanted to compare the effectiveness of initial versus refusal conversion treatments on cooperation, one would want to offer the same amount of money at each stage. The different amounts were chosen for a practical reason—they are equivalent in their expected total cost. Paying money at the initial stage requires sending out money to many more households than at refusal conversion. By equating total cost, this experiment provides a comparison of what a survey designer faces when working with a fixed budget for incentives.

There are other issues related to the use of incentives in an RDD survey. One is understanding the mechanisms around which they work. Use of advance mailing on an RDD survey is not straightforward. Perhaps for this reason, there is only mixed evidence that they increase response rates (Camburn et al. 1995; Brick et al. 1997; Singer et al. 2000; Parsons et al. 2002). To send an advance letter, each telephone number has to be matched against reverse directories to obtain an address. This is successful for only a portion of the numbers. Even for those telephone numbers where there is a match to the directory, some portion may be out of date. Once sending the letter by regular mail, the letter may be thrown away before it is read (Mathiowetz et al. 1994). Even if it is read, there is no guarantee that the person who answers the telephone will be the one who read the letter. The result of this process is that many people who are called do not see the letter the interviewer sends.

Given the complicated sequence of events related to mailing an advance letter described above, it would be useful to know how many respondents actually receive and remember getting the advance mailings. It would be helpful to find out whether the respondent who did not receive the

advance letter knew about the letter or incentive. Furthermore, it would be interesting to find out if the incentive had an effect on the respondent's recollection of the advanced mailing.

2. Experimental Design

The experiment was conducted as part of Round 3 of the National Survey of America's Families (NSAF). The NSAF is a RDD survey funded by a consortium of private foundations in the United States. It is conducted by Westat for the Urban Institute. An important goal of the survey is to assess the impact of recent changes in the administration of government assistance programs for children and the poor.

The NSAF consists of both a screening and extended interview. The screening interview consists of a 3–5 minute battery of questions designed to select the person that should be administered the extended interview. It determines if there are any persons under 65 years old in the household and whether the family is above or below 200 percent of the poverty level. If there is someone in the right age range and the household is sampled (based on poverty status) a respondent for the extended interview is selected. The extended interview is 25–50 minutes long (depending on the type of interview) and covers a wide range of topics, including health, education, child care, income, and receipt of social services. Approximately 42,000 to 45,000 extended interviews are completed in a typical cycle.

The experiment that is discussed below had three conditions, which are shown in table 1. These include the following:

- *Control*—Respondents were sent an advance letter using 1st class mail. If they refused, they were sent a second “refusal conversion” letter in advance to trying a second time. A third refusal conversion attempt was made to those who continued to refuse.
- *\$2 prenotification*—Respondents were sent an advance letter using 1st class mail with a \$2 bill in it. If they refused, they were sent a second “refusal conversion” letter in advance to trying a second time. This refusal conversion letter was sent using USPS priority mail. A third refusal conversion attempt was made to those who continued to refuse.
- *\$5 refusal conversion*—Respondents were sent an advance letter using 1st class mail. If they refused, they were sent a second “refusal conversion” letter in advance to trying a second time. This letter had \$5 in it and was sent using USPS priority mail. A third refusal conversion attempt was made to those who continued to refuse.

To examine the first research question, response rates at the screener were compared across conditions, to address several hypotheses:

Hypothesis 1: Providing an incentive at any stage will increase the response rate.

Hypothesis 2: Sending an incentive at prenotification will be more effective than sending it at refusal conversion.

Hypothesis 3: Sending an incentive will increase the number of respondents who remember receiving the advance material.

Hypotheses 1 and 2 are examined by comparing the response rates across the three conditions. Analysis for hypothesis 1 separately compares the control condition against the \$2 and the \$5 conditions. Analysis for hypothesis 2 compares the \$2 and \$5 condition. Hypothesis 3 is tested using data from questions asked of a sample of respondents about whether they remember receiving the advance materials.

We explore the possibility that the effects of the treatments vary by different population subgroups (research question 2). This is accomplished using geographic data linked to each telephone number. Finally, we assess differences in data quality using two measures from the screening interview (research question 3).

3. Methods

These experiments were conducted at the beginning of Round 3 of the NSAF. The field period for the NSAF was from February to October 2002. Since the experiment was conducted with the initial release groups, most of these cases were finalized between February and April 2002.

All interviewers administering the NSAF during this period participated. This is approximately 300 individuals, once the survey was fully staffed. Interviewers were aware of the different experimental manipulations. The money was mentioned in the introductions to each of these two experimental treatments.

The data reported below are weighted by the initial probability of selecting the telephone number. The weighted data are used to be able to generalize the results to a national population. The significance tests were calculated using *WESVAR 4*©, in conjunction with the JK2 estimation method.

To calculate screener response rates, the following formula was used:

$$SR = (CS)/(CS + R + .63(AM) + .27(NA) + ONR)$$

where *CS* = completed screener, including eligible and ineligible households; *R* = refusals; *AM* = answering machines; *NA* = no answer, and *ONR* = other nonresponse.

4. Response Rates and Awareness by Experimental Condition

Table 2 provides the response rates for the screener by the three different test conditions. The final response rates across conditions were 64.8, 67.9 and 68.2 for the control, \$2 pre-note, and \$5 refusal conversion treatments, respectively. With respect to hypothesis 1, there is a significant difference between the control and the two incentive conditions ($p < .05$; one-tailed test). There is no significant difference between the two incentive conditions (hypothesis 2).

The assumption is that the difference between the incentive and the control is the receipt of the incentive. However, the incentive may also have the effect of drawing attention to the materials that are included in the prenotification package. If a household member opens the mail and sees cash, he or she may be more likely to read the material that is included in the letter. It may convince respondents to read the material. Once doing so, there should be a greater likelihood they will cooperate, regardless of any social obligation related to the payment.

To get a sense of how aware respondents were of the different treatments, the questionnaire included items that asked whether the respondent remembered any of the materials that were sent prior to the call. They were also asked if they told anyone else in the household about these materials and whether anyone else had told them. The items were asked after the respondent had completed the interview.⁵ Table 3 provides these results by the different experimental conditions.

One striking result is the very low rate of recognition for the control condition. Less than one-third of those cooperating remember seeing the letter at the initial call. This is significantly lower than the 60 percent reported for another RDD survey (Parsons et al. 2002).

The variation in this measure across the treatments is consistent with the idea that the incentives serve to bring attention to the prenotification material (hypothesis 3). Sixty-one percent of the respondents in the \$2 prenote condition report remembering the material at the initial stage (table 4). As expected, the rate of recognition for the \$5 refusal conversion treatment is equivalent to the control at the initial call, but increases significantly to 68 percent at refusal conversion.

There also seems to be some communication of the material to other members of the household. When respondents were asked if they had told anyone else in the household about the letter (and incentive), between 40 and 50 percent said they had.⁶ Interestingly, the \$2 produced the most communication in this direction, with the percentage of people telling others being statistically different than the other two treatments at the initial stage. Even at first refusal conversion, the \$2 treatment is nominally the highest, although no differences are statistically significant.

A much lower percentage of the respondents who said they had not gotten the material say that someone else in their household had told them about them the study. The lack of symmetry between these two measures may be indicative of some measurement error. One would have expected that about as many people who said they told someone else would have said that someone else had told them. This may mean that those who are initially responding that they had seen the material may be those who were told about it, rather than seeing it directly.

5. Differences in Data Quality

Two different measures were used to assess data quality (table 5). The first was the amount of missing data to the income item on the screener. This item asked respondents whether their total household income was above or below 200 percent of the poverty level for that household (depends on size and presence of children). As can be seen from these data, there are no significant differences in this measure across the different treatments.

The second indicator was whether the measure of poverty for that household switched between the screener and extended interview. This measure involved comparing the measure of 200 percent of poverty level taken at the screener to the measure collected on the extended interview. The extended interview contains a detailed battery of questions that asks about specific income sources. The income for the entire household is then computed and a variable created for whether

⁵ If the household was deemed not eligible or if the respondent was not selected for an extended interview, these questions were asked right after completing the screener. If the respondent was selected for the extended interview, the questions were asked after they completed the extended interview.

⁶ Only households with more than one person were asked about intra-household communication.

the household was above or below 200 percent of the poverty level. This second measure of poverty status was then compared with the single item on the screener. Higher data quality is indicated by a lower percentage of households switching between these two points in the interview.

As seen from these data, there are no statistically significant differences across the three treatment groups on this measure.

6. Differences across Population Subgroups

As noted in the discussion above, there is some reason to believe that the effect of incentives may be different across subpopulations. One possibility is that incentives will be more effective for low income households. This has been found in other research related to incentives (Singer 2002), although the evidence is mixed. Another possibility is that incentives would be more effective for reluctant respondents. The rationale for the latter is that normal refusal conversion efforts (without incentive) may suffice in convincing those who are marginally reluctant to cooperate. It takes something like an incentive, or the provision of a more tangible benefit, to convince those who are particularly reluctant.

Data from the sampling frame were used to explore whether the incentives were differentially effective over population subgroups. Response rates were tabulated by population groups, which were formed based on the phone numbers' geographic area information. Geographic area is defined by the county or state associated with the number.

Generally, the effects of the treatments do not differ greatly by type of geographic area. Nonetheless, there were a few significant effects that are consistent with the above discussion (table 6). For example, there is a significant difference between the incentive and no-incentive conditions for the low response rate states ($p < .10$; two-tailed test), while no difference exists for the high response rate states. Several of the measures related to socioeconomic status show weak statistical significance. This includes high migration, low owner occupancy, high foreign-born, low employment, and high black population. The family income measure does not show as strong a pattern as one might expect, given the expectation that income is an important correlate of incentives. The incentives were significantly less effective in geographic areas where travel time to work was longer. This relationship is consistent with the time-use research findings that Americans with less free time are more likely to choose to have a day off over an extra day's pay (Robinson and Godbey 1999).

The weak relationships found here may be due to the measures. Counties are large geographic units and there could be quite a bit of variation within these areas. Refining these measures to smaller geographic units may reveal stronger correlations.

7. Discussion

This paper discussed the results of an experiment that explored the use of incentives at different stages of an RDD survey. A primary question addressed was the relative effectiveness of using an incentive at the initial call or during refusal conversion. The results show that providing \$2 at

the initial attempt to complete the screener works about as well with respect to response rates as a \$5 treatment at refusal conversion.

The actual processes incentive affect is not entirely clear. It does seem to be the case that both types of incentives increase the number of persons that report seeing the letter. Of those in the no-incentive group, only about 30 percent of the persons that completed a screener report remembering the letter. This number is essentially doubled when sending \$2 at prenotification or \$5 at refusal conversion. A benefit of the incentive is that it draws attention to the advance material. It is unclear how much of the effect of the incentive adds to the perceived benefits of participating on the survey. This could not be disentangled in this experiment.

There does not seem to be a big difference by incentive groups with respect to whether household members tell other members of the household about the study. While the data presented above are ambiguous, there are no clear patterns across the three experimental groups for whether respondents report telling someone else.

There were no strong differential effects found of the treatments across different subpopulations. There was some suggestion that incentives work better in states with low cooperation rates and in areas that have high migration, low owner occupancy, high foreign-born, low employment, and high black populations. These effects were small at best, generally reaching statistical significance at the 10 percent level. The weakness of these patterns, however, may be a function of the relatively broad geographical areas represented by the measures (counties). Future research should consider refining these measures by narrowing the geographic areas.

Data quality, as measured by the amount of missing data on income and income switching, did not differ across the different treatments.

Overall, the effects of the incentive treatments relative to the no-incentive group were quite small. Providing either type of incentive increased response rates by about 3 to 4 percent. This is considerably lower than that reported by Singer et al. (2000) and Cantor et al. (1998) of 10 percent and 6 percent, respectively. This may be an indication that the public is increasingly becoming resistant to doing surveys, regardless of the use of incentives, at least at the levels tested in this experiment.

It should also be noted that the effects discussed above apply only to that portion of the sample for which an address was found. In the case of this particular survey, that turns out to be approximately 80 percent of the residential households in the sample.

References

Brick, J.M., I. Flores-Cervantes, and D. Cantor. 1999. *1997 NSAF Response Rates and Methods Evaluation*. Methodology Report No. 8.

Camburn, D.P., P.J. Lavrakas, M.P. Battaglia, J.T. Massey, and R.A. Wright. 1995. "Using Advance Respondent Letters in Random Digit Dialing Telephone Surveys." *Proceedings of the Section on Survey Research Methods, American Statistical Association* (969–74). Alexandria, VA: American Statistical Association.

Cantor, D., P. Cunningham, and P. Giambo. 1998. "Testing the Effects of a Pre-Paid Incentive and Express Delivery to Increase Response Rates on a Random Digit Dial Telephone Survey."

Paper presented at the 1998 Annual Meeting of the American Association for Public Opinion Research, St. Louis, MO, May 14–17.

Cantor, D., P. Cunningham, K. Wang, E. Singer, and F. Scheuren. 2002. “An Experiment on the Timing of Incentives and Different Staging Procedures on a Random Digit Dial Survey.” Paper presented at the 2002 Annual Meeting of the American Association for Public Opinion Research, St. Petersburg, FL, May 16–19.

Dillman, D.A. 2000. *Mail and Internet Surveys: The Tailored Design Method*. New York: John Wiley.

Mathiowetz, N.A., M.P. Couper, and E. Singer. 1994. “Where Does All the Mail Go? Mail Receipt and Handling in U.S. Households.” Unpublished paper. U.S. Bureau of the Census.

Parsons, J., L. Owens, and W. Skogan. 2002. “Using Advance Letters in RDD Surveys: Results of Two Experiments.” *Survey Research* 33(1): 1–2. Newsletter from the Survey Research Laboratory, College of Urban Planning and Public Affairs, University of Illinois, Chicago.

Robinson, John P., and Geoffrey Godbey. 1999. *Time for Life: The Surprising Way Americans Use Their Time*. Pennsylvania: Penn State University.

Singer, E. 2002. “The Use of Incentives to Reduce Nonresponse in Household Surveys.” In *Survey Nonresponse*, edited by R. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (163–78). New York: John Wiley.

Singer, E., J. Van Hoewyk, and M.P. Maher. 2000. “Experiments with Incentives on Telephone Surveys.” *Public Opinion Quarterly* 64:171–88.

Table 1. Experimental Design

	Condition		
	Control	\$2 prenotification	\$5 refusal conversion
Initial attempt	Letter, 1st class	Letter, 1st class; prepay \$2	Letter, 1st class
Refusal conversion	Letter, 1st class	Letter, priority	Letter, priority; advance \$5

Table 2. Response Rates by Experimental Treatment and Stage

	Control	\$2 prenotification	\$5 refusal conversion
Initial attempt	45.9	51.2 ^{+#}	43.0 [#]
After refusal conversion	64.8 ⁺	67.9 ⁺	68.2 [*]
(Unweighted <i>N</i>)	(1,814)	(16,499)	(3,665)

⁺ Control versus \$2 significant at $p < .05$ (one-tailed test)

^{*} Control versus \$5 significant at $p < .05$ (one-tailed test)

[#] \$2 versus \$5 significant at $p < .05$ (one-tailed test)

Table 3. Percent Remembering Letter, by Experimental Group

	Control	\$2 prenotification	\$5 refusal conversion
With addresses			
Initial			
(unweighted <i>N</i>)	27.4 ⁺	61.0 ^{+#}	33.6 [#]
At first refusal conversion	(227)	(1966)	(425)
(unweighted <i>N</i>)	35.2 ⁺	43.2 [#]	67.6 ^{+#}
Total after all refusal conversion	(65)	(454)	(197)
(unweighted <i>N</i>)	29.2	56.2	44.9
(unweighted <i>N</i>)	(309)	(2595)	(666)
Without addresses			
Total after all refusal conversion	9.5	2.1	10.8
(unweighted <i>N</i>)	(19)	(157)	(38)

⁺ Control versus \$2 significant at $p < .001$ (one-tailed test)

^{*} Control versus \$5 significant at $p < .01$ (one-tailed test)

[#] \$2 versus \$5 significant at $p < .001$ (two-tailed test)

Table 4. Percent Who Communicated Receipt of Materials, by Experimental Group

	Control	\$2 prenotification	\$5 refusal conversion
Tell others in the household?	45.5	52.9	34.8 [#]
(Unweighted <i>N</i>)	(78)	(1,317)	(241)
Others tell you?	2.1	4.0	4.4
(Unweighted <i>N</i>)	(192)	(1,016)	(357)

[#] \$2 versus \$5 significant at $p < .01$ (one-tailed test)

Table 5. Indicators of Data Quality

	Control	\$2 prenotification	\$5 refusal conversion
Missing on income			
% Don't know	3.6	2.9	3.4
% Refused	2.5	2.7	2.9
(Unweighted <i>N</i>)	(746)	(6952)	(1553)
% switched poverty status (total)	12.2	11.3	10.5
(Unweighted <i>N</i>)	(494)	(4673)	(1045)

Table 6. Final Response Rate by Experimental Condition and Selected Area Characteristics

	Control (1)	\$2 prenotification (2)	\$5 refusal (3)
High response rate state	64.26	66.45	65.84
Low response rate state	55.30 ⁺ *	58.56 ⁺	59.85 [*]
Inside central city of MSA	58.02	62.08	60.64
Outside a central city of MSA but in central city's county	59.92	61.27	64.89
Inside a suburban county of an MSA	60.61	62.29	62.64
In an MSA that has no central city	53.63	57.52	59.40
Non-MSA	68.41	71.02	69.95
High travel to work	61.14	60.51	60.34
Low travel to work	60.48 ^{#x}	68.87 [#]	69.41 ⁺
High migration	59.09 ^{#*}	64.12 [#]	63.24 [*]
Low migration	64.68	62.27	64.20
High owner occupancy	65.33	66.10	68.85
Low owner occupancy	59.02 ⁺	62.47 ⁺	61.57
High foreign-born	57.93 ⁺	61.16 ⁺	61.38
Low foreign-born	67.43	68.92	69.24
High employment	62.71	63.72	64.94
Low employment	57.94 ⁺	63.15 ⁺	61.55
High family income	60.39	62.25	62.71
Low family income	63.23	68.05	66.83
High black population	56.20 ⁺	60.84 ⁺	59.93
Low black population	65.49	66.12	67.06
High household size	59.24	61.80	60.47
Low household size	62.61	65.25	66.69
High presence of children	60.31	64.06	63.55
Low presence of children	61.50	62.94	63.56

⁺ Control versus \$2 significant at $p < .10$ (two-tailed test)

^{*} Control versus \$5 significant at $p < .10$ (two-tailed test)

[#] Control versus \$2 significant at $p < .01$ (two-tailed test)

^x Control versus \$5 significant at $p < .01$ (two-tailed test)

Comparing Promised and Prepaid Incentives for an Extended Interview on a Random Digit Dial Survey

David Cantor, Westat; Kevin Wang and Natalie Abi-Habib, The Urban Institute

1. Introduction

This paper reports on a test of the use of incentives on a random digit dial (RDD) survey. While there has been quite a bit of research on incentives across all modes of interviewing, very little has been conducted for RDD surveys. Of the research that has been done, most relates to the use of incentives at the initial contact with the household. A set of experiments described by Singer et al. (2000) present evidence that incentives are effective in an RDD context at the initial stage. They find that a prepayment of \$5 did significantly improve response rates in a series of experiments involving the Survey of Consumer Attitudes (SCA). A promised incentive did not increase response rates. Similarly, Cantor et al. (1997, 1998; Brick et al. 1999) found that small, prepaid incentives work at both the initial and refusal conversion stages of the process. Promised incentives were not found to work.

It is common on a RDD survey to have both screening and extended levels to the survey process. A screening survey is administered to a general population to find units with specific characteristics selected. A longer, extended interview is then completed with the selected respondent. The issues of concern in this paper are the effectiveness of incentives at the extended level. Specifically, this paper examines four questions: (1) How does the effectiveness of a promised incentive at the extended interview compare with not using any incentive at all? (2) How does the effectiveness of a promised incentive at the initial contact compare with an advance incentive at refusal conversion? (3) Is the effectiveness of an incentive for an extended interview affected by the type of incentive offered at the screener? (4) Are there differences in data quality and/or the characteristics of respondents by the type of incentive offered at the extended interview?

2. Promised Incentives after Screening Households

The literature on incentives finds that prepaid incentives are more effective than promised incentives (Singer et al. 1999). According to social exchange theory, a prepaid incentive works because it increases the social benefits to the respondent (Dillman 2000, 15–22). It legitimizes the survey by showing the respondent the researcher is willing to provide a benefit, no matter how small, before the respondent has completed any tasks. This invokes a “social obligation” on the respondent. Once making the incentive contingent on completing a task, as with a promised incentive, the exchange shifts from a social to an economic one. Once viewed as a purely economic exchange, the monetary rewards may not measure up to the perceived burden of the task.

One of the difficulties related to using promised incentives on an RDD survey is communication. Many of the refusals on an RDD survey occur within the first 10–15 seconds of the interaction. There isn’t much time for the interviewer to communicate details either about the survey or how the incentive may be related to the associated task. Promises of money at this point may actually

have a negative effect on cooperation, because the offer may be confused with offers of money that some telemarketers make to get respondent's attention.

These communication problems should be less of an issue when asking for cooperation on an extended interview. At this point in the process, a screening interview has already been completed and the respondent is actively listening to what the interviewer is saying. The respondent is likely to have a bit more confidence in the credibility of the interviewer and the study.

Given the above, we pose two hypotheses:

Hypothesis 1: A promise of money for the extended interview will significantly increase response rates relative to not promising anything at all.

Hypothesis 2: A promise of money for the extended interview will increase response rates relative to sending a smaller amount of money in advance of the refusal conversion call.

Hypothesis 2 was tested because sending money at refusal conversion is an increasingly common practice for survey designers. Cantor et al. (1998) found that when used at the screener, this method produces response rates comparable to sending money in advance of the call. Providing the incentive at refusal conversion is consistent with Dillman's (2000) idea of creating an increasing sense of reward for participating in the survey. When viewed in the context of all the contacts made with the household, offering a refusal payment recognizes the respondent's initial reluctance to participate in the study. Respondents may appreciate the persistence of the interviewer and the idea that someone places such a high value on their views and time. This would predict that offering an incentive at this stage would be quite effective.

On the other hand, one might question offering an incentive at this later stage because it may change the exchange from a social to an economic one. This shift may occur because the sequence of contacts resembles a bargaining exchange. The respondent first refuses to participate, which leads to a monetary offer to cooperate. As noted above, once viewed as an economic, rather than social, exchange, the monetary rewards may not measure up to the perceived burden of the task. Related to this problem is that respondents may wonder why an incentive was not provided at the initial contact. They may become suspicious about the motives of the survey administrators.

3. Effect of Screener Incentives on the Extended Interview

There is very little research on the use of incentives at multiple points in the survey process. Similarly, for a cross-sectional survey, one might provide an incentive when screening for eligible respondents and a second incentive when interviewing the individual selected for the survey. While there is quite a bit of research on the best way to use incentives for the initial interview (e.g., screening interview for cross sectional survey), there is very little guidance on the use of incentives over multiple contacts and situations. For example, it is unclear whether it is better to provide a large, one-time incentive at the first wave of a panel or smaller incentives at each wave. There is a similar problem in many RDD surveys, which require both a screening and an extended interview.

The third research question listed above concerns the interaction of the type of incentive at the screener and the extended interview. The specific concern is whether the sequencing of the screener incentive affects response rates at the extended interview. “Sequencing” refers to whether the incentive is offered when the screener is initially attempted or at refusal conversion. The hypothesis tested below is:

Hypothesis 3: A screener incentive offered at the initial contact will increase the response rate at the extended interview compared with screener incentives offered at refusal conversion.

As noted above, one argument against the use of incentives at refusal conversion is that it communicates an economic, rather than a social, exchange. Once doing this at the screener, there may be a tendency for respondents to view all subsequent contacts with the survey in this light.

4. Incentives and Response Distributions

Incentives may affect response distributions in a number of different ways. One is by changing the amount of missing data that occurs on the survey. Incentives may attract respondents who are less willing to participate in the survey and are more likely to provide poor quality data (e.g., by answering questions incorrectly or with too little thought, or refusing to answer questions). Another possibility is that incentives serve to motivate respondents and as a result, they provide better quality data (e.g., more carefully thought-out answers, lower item nonresponse rates).

Similarly, incentives may be more attractive to certain kinds of respondents. For example, one hypothesis is that they will be most attractive to respondents in low income groups (Singer 2002).

The final sections of the analysis explore these possibilities by analyzing the response distributions of key indicators (e.g., missing data on income, demographic characteristics) by the type of incentive that is offered.

5. Experimental Design

The experiment was conducted as part of Round 3 of the National Survey of America’s Families (NSAF). The NSAF is a RDD survey funded by a consortium of private foundations in the United States. It was conducted by Westat for the Urban Institute. An important goal of the survey is to assess the impact of recent changes in the administration of a number of assistance programs for children and the poor.

The NSAF consists of both a screening and an extended interview. The screening interview consists of a 3–5 minute battery of questions that is designed to select the person that should be administered the extended interview. This involves determining if there are any persons under 65 years old in the household and whether or not the family is above or below 200 percent of the poverty level. If someone is in the right age range and the household is sampled (based on poverty status), a respondent for the extended interview is selected. The extended interview is 25–50 minutes long (depending on the type of interview) and covers a wide range of topics, including health, education, child care, income, and receipt of social services. Approximately 42,000 to 45,000 extended interviews are completed in a typical cycle.

The design of the experiment is shown in table 1. There were two experimental factors. The first was a *screener incentive*, which included sending a \$2 incentive along with a letter before making the first call, a \$5 incentive along with a letter before calling to convert refusals, or a letter without an incentive. The second factor was the extended incentive, including promising money when first requesting an extended interview, and sending \$5 before trying to convert refusals and promising an additional \$20 if the interview is completed.

Crossing these two factors yields four experimental groups, as shown in the first 4 columns of table 1. The “promise” condition for the extended interview had two levels of incentives. The study was interested in offering an extra incentive to populations that were of special interest or had shown reluctance to participate in the past. Those individuals that did not report their income on the screener and those individuals that were not located in one of the 13 states with an oversample were offered \$20 (approximately 30 percent of the sample). All other persons in the sample were offered \$10 (approximately 70 percent of the sample).

The fifth group (column) shown in table 1 is the “no treatment” group, which did not provide an incentive to any household at either the screener or extended level.

6. Methods

These experiments were conducted at the beginning of Round 3 of the NSAF. The field period for the NSAF lasted from February to October 2002. Since the experiment was conducted with the initial release groups, most of these cases were finalized between February and July 2002.

All interviewers administering the NSAF during this period participated. This is approximately 300 individuals, once the survey was fully staffed. Interviewers were aware of the different experimental manipulations.

The NSAF produces estimates for two different population groups. The first are families with at least one child age 0–17. To collect these data, the person selected to do the extended interview is the person who knows the most about the child that was sampled during the screener (most knowledgeable adult, or MKA). The second population are all adults who are of working age (18–64). The respondent for this group is selected in several different ways. One method is to administer the extended interview to a randomly selected adult who is living in a household where there are no children present. Adults are also selected within households where there are children. The results reported below tabulate the results for the MKA and for the adults living in households without children (referred to as “adult-only” households).

Two rounds of refusal conversion were completed for most households. The results are weighted by the initial probability of selecting the telephone number, a nonresponse adjustment done at the screener level, and the probability of selecting the household once the correct information was obtained. The latter included, for example, accounting for any oversampling that was done for those under 200 percent of the poverty level. The nonresponse adjustment at the extended level was not included in the weights applied below. These weights do not account for the probability of selecting a particular respondent within the household. The weighted data are used to be able to generalize the results to a national population. The significance tests were calculated using *WESVAR 4*©, in conjunction with the JK2 estimation method.

To calculate response rates, the following formula was used:

$$SR = (CI)/(CI + R + ONR)$$

Where *CI* = completed interviewer, *R* = refusals, and *ONR* = other nonresponse (includes non-contacts, broken appointments, answering machines, field period ending, language problems and other types of nonresponse).

7. Results

The results are discussed according to the research questions and hypotheses discussed above.

Does a promised incentive work at the extended level?

Initially, our interest was in knowing whether the promise of money at the extended interview significantly increased the response rates relative to promising nothing at all (hypothesis 1). This was addressed using the data displayed in table 2, which provide the response rates once collapsing across the experimental groups with a common treatment at the extended interview. The first row is for interviews with MKAs (respondents reporting for a sampled child) and the second row is for a randomly selected adult in an adult-only household.⁷ The first column is the promise of money, which adds together groups 1 and 2 in table 1. The second column is the extended treatment using a \$5 refusal conversion payment with the \$20 promise for completing the survey (groups 3 and 4 in table 1). The third column is for the group that was not offered any incentive at all (column 5 of table 1).

Strictly speaking, these groups are not entirely equivalent because the “no treatment” group did not have a screener incentive, while the other two each had some type of incentive treatment at the screener. As will be shown below, this may affect how the extended incentive is received in the household. Nonetheless, this comparison does provide an indication of whether a promise of money at the extended level has the potential to increase response rates.

For the MKAs, the promise of money is significantly different from not providing any incentive at all (84.9 versus 75.8 percent; $p < .05$; two-tailed test). This is not the case for the interviews with respondents in the adult-only households. In this case, the promised incentives are about the same level as not promising anything at all (82.3 versus 85.4 percent). A similar pattern is apparent for the other incentive treatment (\$5 at refusal conversion with promise of \$20). This is also significantly different for the MKAs from the no treatment group (82.2 versus 75.8 percent; $p < .10$; one-tailed test). It is not significantly different for the adult-only households.

The second hypothesis posed above was whether the extended treatments differed from one another. This can be tested by comparing the first two columns of table 2. The effects of these different extended incentive schemes do not differ by the type of respondent. For the MKAs, the rates are 84.9 versus 82.2 percent and for adult-only respondents the rates are 82.3 and 78.9 percent. While the promised incentive is higher for both types of respondents, none of the differences are big enough to reach statistical significance.

⁷ Analysis not discussed here found that there was not a significant difference between offering \$10 or \$20 at the extended level. Consequently, the results below aggregate together cases that were promised either \$10 or \$20 for the extended interview.

Does the screener incentive affect extended interview response rates?

Hypothesis 3 above concerns whether the staging of the screener incentive interacts with incentives provided at the extended level. The initial hypothesis was that providing \$2 to all households will have a bigger positive effect on subsequent procedures at the extended than using screener incentives at refusal conversion. The primary rationale being that the \$2 treatment is prepaid and establishes a clear social exchange before the initial contact with the survey, while the \$5 may shift the motivation from a social to an economic one.

Table 3 provides support for this hypothesis. These data are the response rates disaggregated by both the screener treatments and extended treatments. These columns correspond to the five experimental groups shown in table 1. From these data, there does seem to be an effect of the screener treatment on extended response rates. It is strongest for the adult-only households, where the rate for the promised incentive is 10 points higher when the \$2 screener treatment was used compared with the \$5 refusal conversion treatment (83.6 versus 73.0 percent; $p < .05$; two-tailed test). A similar difference appears for the \$5 refusal conversion extended incentive treatment, where the difference is also around 10 points (83.2 versus 73.6 percent; $p < .10$; one-tailed test). These patterns carry over to the MKAs but are much smaller and not statistically significant (85.3 versus 82.4 percent and 83.7 versus 80.9 percent).

Do incentives affect the amount of missing data?

The amount of missing data was estimated for key items from across the extended interview treatment conditions (data not shown). For MKA interviews, there is a tendency toward higher levels of item nonresponse on earnings and income items for those being offered incentives relative to those not offered any incentives. This does not appear to be the case with the adult-only respondents. However, analysis that controls for other factors associated with item nonresponse should be carried out before concluding that the use of incentives on the extended interview increases levels of item nonresponse. It may be the case, for example, that respondents in the no-treatment conditions answered different (e.g., fewer) income items, which may have led to a lower rate of missing data.

Do incentives at the extended interview affect respondent characteristics?

Another concern related to incentives is that it affects the types of respondents who agree to complete the survey. Analysis was conducted that examined key characteristics of respondents by each extended treatment group. In general, for demographic items that are used in the NSAF population weighting adjustments (home ownership, race/ethnicity, age, education), there were very few significant differences across the three treatment conditions.

A similar analysis was completed for key survey items across the three treatment conditions. Most of the differences in estimates across the treatment conditions were not statistically significant. For MKA interviews, there is some evidence that respondents in the incentive conditions tend toward higher socioeconomic status, especially with respect to employment. One possible explanation for this pattern is that for higher-income respondents, incentives may compensate for a lack of interest in the subject matter of a survey that focuses primarily on the well-being of low-income families. This pattern does not occur for respondents in households without children.

8. Discussion

This analysis was structured around three questions. The first compared the effectiveness of a promised incentive at the extended interview. The analysis above provides evidence that promises of money at this level work for certain kinds of respondents. Significant effects were found for the MKAs. No effects were found for the adult-only group. These two groups of respondents differ demographically. For example, MKAs are more likely to be female and married. The survey procedures also treat these two groups differently. There is more discretion on the part of the screener respondent when selecting the MKA, since it is based on the respondent's judgment about who can answer questions about the sampled child. The adult-only respondent is selected at random from a list of persons living in the household. As a consequence, a higher proportion of the MKAs are also the screener respondent. One might expect that communication about the survey and the conditions surrounding participation could be different for those who are screener respondents and those who are asked to participate once the screener is completed by someone else.

A second possibility is that the weights used in the analysis did not fully account for the respondent's probability of selection. Each respondent was assigned a weight accounting for the households chance of selection. However, this weight did not account for the chance of selection within the household. Those in larger households should have relatively higher weights than those in smaller households. The weights in the current analysis do not reflect this. Future analysis should recompute the above response rates using the correct weights to assess whether this is related to the differences across adult-only and MKA respondents.

The second question was concerned with whether the effectiveness of an initial promised incentive at the extended interview is different than an advance incentive offered at refusal conversion. When these two treatments were compared, no differences were found with respect to the effects on response rates. This was true for both MKA and adult-only interviews. On its face, this result is similar to that found for research at the screener level, where the use of incentives at either the initial or refusal conversion stages yield approximately the same response rates.

The third question was whether the treatment at the screener affected results at the extended level. The experiment tested whether the sequencing of the screener incentive (initial versus refusal conversion) influenced the effects of the incentives at the extended level. The above analysis provides evidence that this was true for at least the adult-only respondents. Use of a refusal conversion treatment at the screener seemed to depress the extended interview rates, regardless of the type of incentive offered. The worst combination seemed to be the use of refusal conversion treatments at both the screener and extended levels. A similar pattern was found for the MKA households, but the differences were not as large or statistically significant.

These last results suggest that application of incentives at early stages of a survey do have effects at later stages. They would further suggest that use of refusal conversion payments may be less effective from this perspective than a prepaid incentive at the initial contact. It is unclear why this may be the case. It may be because the prepaid incentive reaches all sample members and, thus, sets up a social exchange that is viewed favorably by many in the household. The refusal conversion payment reaches fewer people. Alternatively, it may be because the screener

conversion payment makes it appear as if the survey is trying to buy the respondent's cooperation. Once doing this, the motivation to cooperate at later stages may go down.

The no-incentive condition had a lower prevalence of missing data than the two incentive treatments. On its face, this implies that providing incentives may decrease motivation to respond. No strong differences were found in the response distributions across the different treatments.

References

Brick, J.M., I. Flores-Cervantes, and D. Cantor. 1999. *1997 NSAF Response Rates and Methods Evaluation*. Methodology Report No. 8.

Cantor, D., B. Allen, P. Cunningham, J.M. Brick, R. Slobasky, P. Giambo, and G. Kenny. 1997. "Promised Incentives on a Random Digit Dial Survey." In *Nonresponse in Survey Research*, edited by A. Koch and R. Porst (219–28). Mannheim, Germany: ZUMA.

Cantor, D., P. Cunningham, and P. Giambo. 1998. "Testing the Effects of a Pre-Paid Incentive and Express Delivery to Increase Response Rates on a Random Digit Dial Telephone Survey." Paper presented at the 1998 Annual Meeting of the American Association for Public Opinion Research, St. Louis, MO, May 14–17.

Dillman, D.A. 2000. *Mail and Internet Surveys: The Tailored Design Method*. New York: John Wiley.

Singer, E. 2002. "The Use of Incentives to Reduce Nonresponse in Household Surveys." In *Survey Nonresponse*, edited by R. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (163–78). New York: John Wiley.

Singer, E., J. Van Hoewyk, and M.P. Maher. 2000. "Experiments with Incentives on Telephone Surveys." *Public Opinion Quarterly* 64:171–88.

Singer, E., J. Van Hoewyk, N. Gebler, T. Raghunathan, and K. McGonagle. 1999. "The Effect of Incentives on Response Rates in Interviewer-Mediated Surveys." *Journal of Official Statistics* 15:217–30.

Table 1. Experimental Design

Treatments	Condition number				
	1	2	3	4	5
<u>Screening Interview</u>					
Pre-pay \$2	x		x		
Advance \$5 before refusal conversion		x		x	
No incentive					x
<u>Extended Interview</u>					
Promise \$10 to 13 sites; promise \$20 to balance of nation ⁺	x	x			
Advance \$5 at refusal conversion and promise \$20 when completing			x	x	
No incentive					x

Notes: Only includes households where an address could be found. All treatments include sending a letter in advance to making the initial call.

⁺ = \$10 offer was made to households that provided an answer to the screening income item and who are in the 13 targeted states. \$20 offer was made to households that were not in the targeted 13 states or were missing on the income screening item.

Table 2. Extended Interview Response Rates, by Incentive at the Extended

	Promise money at extended interview	Prepay \$5/promise \$20 at extended	No incentive at extended	Significance
MKA households				
Response rate (unweighted <i>n</i>)	84.9 2245	82.2 649	75.8 300	1 versus 3 ^{**} ; 2 versus 3 [#]
Adult-only households				
Response rate (unweighted <i>n</i>)	82.3 898	78.9 284	85.4 134	ns

⁺ Includes sample where an address was found

* Significant at $p < .10$ (two-tailed test); ** Significant at $p < .05$; # Significant at $p < .10$ (one-tailed test).

Table 3. Extended Interview Response Rates, by Experimental Group

	<u>Promise Money at Extended Interview</u>		<u>Prepay \$5/Promise \$20 at Extended</u>		<u>No Incentive at Extended</u>	<u>Significance</u>
	<u>Screeners Treatment</u>		<u>Screeners Treatment</u>		<u>Screeners Treatment</u>	
	\$2 at initial	\$5 at conversion	\$2 at initial	\$5 at conversion	No incentive	
MKA households						
	1920					
Response rate (unweighted <i>n</i>)	85.3 325	82.4 323	83.7 326	80.9 300	75.8	1 versus 5 ^{**} ; 3 versus 5 [*]
Adult-only households						
Response rate (unweighted <i>n</i>)	83.6 773	73.1 125	83.2 146	73.6 138	85.4 134	1 versus 2 [*] ; 1 versus 4 [*] ; 3 versus 4 [#] ; 2 versus 3 [*] ; 2 versus 5 [*]

⁺ Includes sample where an address was found

^{*} Significant at $p < .10$ (two-tailed test); ^{**} significant at $p < .05$ (two-tailed test); [#] significant at $p < .10$ (one-tailed test)

2. PAPERS PRESENTED AT THE AMERICAN STATISTICAL ASSOCIATION'S ANNUAL JOINT STATISTICAL MEETINGS

Overview

Chapter 2 consists of four NSAF methodology papers that have been presented at the annual Joint Statistical Meetings (JSM) and have not been published elsewhere besides in the survey research methods section of the conference proceedings. While most of these papers can be found in the JSM proceedings, the versions included in this report may vary owing to the page size and time deadlines associated with papers submitted for proceedings publication.

Only the third paper in this report uses data from the 2002 NSAF (“Sampling Refusals: Why, When, and How Much?”). The other papers in this chapter use data from the earlier rounds of NSAF, but are included in this report because they were completed after the release of the 1999 collection of papers methodology report.

Using Paradata to Examine the Effects of Interviewer Characteristics on Survey Response and Data Quality

Adam Safir, Tamara Black, and Rebecca Steinbach

1. Introduction

This research began with the premise that both survey organizations and data users can benefit from the release of paradata. The release of paradata, or administrative data, to accompany other public-use survey data make possible a more comprehensive and independent evaluation of nonsampling error in survey estimates of interest, and is in keeping with Deming's idea of a "full-systems thinking" approach to data quality (Scheuren 2001). We attempt to demonstrate the feasibility of this approach through a practical example using data from the National Survey of America's Families (NSAF), with a particular emphasis on interviewer-related error.

While macro paradata represent global process summaries (e.g., sampling error or response rates), micro paradata (hereafter to be referred to as simply paradata) provide process details about specific interviews on a case-by-case basis, such as how many attempts were required to contact a sampled unit or the tenure of the interviewer who completed the case. In general, survey organizations do not release an exhaustive set of paradata items. This may be for any number of reasons, including legitimate confidentiality concerns, lack of researcher interest, or inability by data producers to see any clear value added (Scheuren 2001).

The two complimentary research objectives, therefore, were to better understand the costs and benefits of releasing paradata along with other survey data on public use files, and, endemic to this process, to examine the relationship between known interviewer characteristics and data quality. Using administrative data from the NSAF, the analysis described in this paper grouped telephone interviewers according to predetermined characteristics and compared measurements on survey items of interest.

2. Survey Background

The NSAF is a survey of the economic, health, and social characteristics of children, adults under the age of 65, and their families. Two rounds of interviews have been conducted. The first round was conducted from February 1997 through November 1997, and the second round from February 1999 through October 1999. Each round of interviews yielded information on over 40,000 households and 100,000 persons. Westat conducted the data collection for the NSAF.

The survey is representative of the civilian, noninstitutionalized population under age 65, and features an oversample of low-income households with children at the state level. Large representative samples of households were taken in each of 13 targeted states and the balance of the nation. The 13 states were Alabama, California, Colorado, Florida, Massachusetts, Michigan, Minnesota, Mississippi, New Jersey, New York, Texas, Washington, and Wisconsin. These 13 states represent over half of the U.S. population and reflect a broad array of government programs, fiscal capacity, and approaches to child well-being. Collectively, the survey estimates form what we believe to be a sound baseline from which many of the changes brought about during the period of devolution can be measured and assessed (Converse et al. 2001).

3. Interviewer-Related Error

In general, error can be defined as the difference between a survey answer and the true value of what the researcher is interested in measuring. The degree of interviewer-related error (only one component of total survey error) in the measurement process is related to the extent to which interviewers can be associated with the answers they obtain (Fowler and Mangione 1990).

Unique interviewer characteristics may affect survey responses for a number of reasons. For example, interviewers have a range of skill sets—some excel at gaining cooperation and overcoming the objections of reluctant respondents, while others are better able to administer a standardized interview. Second, certain interviewer characteristics may alter the context or meaning of questions. Finally, interviewer characteristics may affect the quality of the relationship between the interviewer and respondent (Fowler and Mangione 1990). In other words, interviewers may alter the delivery of the survey script depending on factors related to tenure, length of shift, and/or perception of the survey.

Interviewer error is particularly important in telephone surveys, where a small number of interviewers may complete a large number of interviews (Singer et al. 1983). To the extent that there exists a systematic variation in responses directly attributable to the interviewer, paradata represent a useful tool for ascertaining the degree, and in some cases, the ignorability, of interview-related error.

4. Methods

4.1 Analytic Objectives

Our intention was to accomplish the first research objective, assessing the utility of paradata, through the pursuit of the second objective: analyzing the relationship between known interviewer characteristics culled from paradata on the public-use files and items reflecting various components of survey response. The three hypotheses stemming from this second objective can be described as follows:

Skill: It was hypothesized that the relative skill of the interviewer might influence the interviewer-respondent relationship in such a way as to produce a noticeable effect on survey statistics.

Tenure: It was hypothesized that interviewers who had worked on the survey in the past, and were therefore more familiar with the questionnaire and subject matter, might have developed habits that would result in a perceptible and identifiable impact on the interview.

Experience: It was hypothesized that the effects of current accumulated experience on the survey would be evidenced in the survey statistics.

4.2 Variable Descriptions

Independent Variables. Two broad categories of interviewing skill are effectiveness in gaining cooperation and ability in asking survey items and recording responses accurately. While a more

complete analysis of interviewing skill would include paradata variables reflecting as many of these measures as possible (e.g., cooperation rate, percent of monitored questions asked exactly as worded, accuracy in data entry, etc.), the sole measure of skill contained in the administrative data was cooperation rate quartile. Therefore, two paradata variables were selected of this type: SCCOCOOP, which grouped interviewers into screener cooperation rate quartiles, and EXCOCOOP, which grouped interviewers into extended interview cooperation rate quartiles.

For tenure, the variable EXCOWRKN, which indicated whether the interviewer had worked on the survey in the first round of data collection, was selected. This variable was meant to provide a measure of the interviewer's overall familiarity with survey.

Finally, to measure current accumulated experience, the variables INTCNT and TOTINT were selected. For each record, INTCNT indicated the number of cases completed by the interviewer who completed that particular interview. The value of TOTINT reflected the total number of interviews completed by the interviewer who completed a specific case.

Dependent Variables. The two types of dependent variables selected for the analysis were specific questionnaire items and more general (or global) survey response measures.

To analyze the effect of interviewer characteristics on response to questionnaire items, questions were classified into four groups: factual/nonsensitive, factual/sensitive, subjective/nonsensitive, and subjective/sensitive. In the context of the NSAF, the term "sensitive" refers mainly to questions that the respondent might consider sensitive relative to the other questions in the survey. The term "subjective" is used to distinguish between personal items, such as opinion questions, and more factual items, such as questions about welfare receipt or family income.

Using this classification, we expected to see a loose hierarchy of effects. That is, the subjective/sensitive measures were expected to be the most likely to be influenced by characteristics of the interviewer, and the factual/sensitive and subjective/nonsensitive variables to a lesser extent. The factual/nonsensitive variables were selected primarily for control purposes.

The analysis also examined three more global measures of survey response and data quality. These included interviewer rates of income switching, average interview length, and item nonresponse. Rates of income switching, along with average interview length and item nonresponse rates on selected variables were determined by interviewer for all cases included in the final analysis file.

4.3 Analysis

Logistic regression was employed to test the hypotheses regarding skill, tenure, and response to groups of questionnaire items (figure 1). The logit model included the screener cooperation rate quartile of the interviewer who completed the case (SCQ), the extended cooperation rate quartile of the interviewer who completed the case (ECQ), and the tenure of the interviewer who completed the case (WR1). Dummy variables were created for the cooperation rate quartile variables in the model. Respondent race (RR) and metropolitan status (MET) were entered into the model as a way of controlling for effects related to sample composition.

Figure 1. Logit Model

$$\text{logit}(Y) = \alpha + \beta(\text{SCQ}) + \beta(\text{ECQ}) + \beta(\text{WR1}) + \beta(\text{RR}) + \beta(\text{MET})$$

Chi-square tests of independence were used to test the survey measures hypothesis on experience. Linear regression was employed to relate skill, tenure, and experience (total interviews completed, or “TI”) to the more global measures (figure 2).

Figure 2. OLS Model

$$Y = \alpha + \beta(\text{SCQ}) + \beta(\text{ECQ}) + \beta(\text{WR1}) + \beta(\text{TI})$$

Although we expected the screener and extended cooperation rate quartile variables to be highly correlated, in fact they were not. And as can be seen in figure 3, the distribution of interviewers across screener and extended cooperation rate quartile does not display the dominant clustering along the diagonal that one might expect.

Figure 3. Cooperation Rate Quartile Comparison

Screener	Extended				Total
	1st	2nd	3rd	4th	
1st	28.3	32.1	21.7	17.9	100
2nd	23.7	30.7	24.6	21.1	100
3rd	22.2	27.8	34.1	15.9	100
4th	7.5	25.3	18.1	48.9	100

While close to 50 percent of the 4th screener cooperation rate quartile interviewers can be found in the comparable extended cooperation rate quartile, just 28 percent of the 1st screener cooperation rate quartile interviewers are in the 1st extended cooperation rate quartile.

Finally, it is important to note that in analyzing the regression results, there was less interest in the summary statistics for the model’s explanatory power as there was in examining the behavior of the coefficients to shed light on whether there was a nonrandom, statistically significant difference across interviewer classifications that might contribute bias to the sample results.

4.4 Controls

Under ideal survey conditions, interviewer assignment is totally random and interpenetrated; however, in practice, this is rarely the case. Some interviewers work exclusively the day shift, others the evening shift, and still others, designated as refusal conversion specialists, may be assigned only the most difficult cases. To account for the nonrandom assignment of cases to interviewers, the analytic data set was limited to those completed cases that had never refused and that had been started and completed by the same interviewer. This latter component was intended to control for completed break-offs, in which a different interviewer completed a case that had been started by another interviewer. To control for additional possible differences related to sample composition, the data set was also limited to households with children. The final sample size for the analysis file was 12,711 cases.

5. Findings

5.1 Questionnaire Items (Skill and Tenure)

Subjective/nonsensitive. The first group of variables tested in the logistic regression model were the subjective/nonsensitive variables. These were mainly variables measuring opinions about issues such as welfare, parenthood, and childbearing. As can be seen in table 1, the final analysis revealed very little association between characteristics of the interviewer and question response. Although some of the regression coefficients showed significance, no clear patterns emerged.

Subjective/sensitive. The subjective/sensitive variables included indices measuring parental aggravation, behavioral problems of children, children's school engagement, and mental health of parents. For these variables, the regression model was predicting that the response would be what one might consider "nonsensitive," that is, no parental aggravation, no behavioral problems, and so on.

The regression coefficients for these variables revealed more interesting trends. Table 1 displays the observed patterns in the screener cooperation rate coefficients. Although the results for some items suggested that interviewers in higher cooperation rate quartiles obtained more sensitive responses, the pattern was considered inclusive because of the absence of significant results. On other items, somewhat more curious patterns were observed. For example, in some cases the first and fourth cooperation rate quartile interviewers appeared more similar in relative size of the coefficient, or the coefficients for the second and fourth cooperation rate quartile interviewers appeared more similar, but a stepped, ordered relationship between cooperation rate quartile and item response was not evident. The regression coefficients for the extended cooperation rate mirrored these results.

Factual/sensitive. The factual/sensitive measures were made up of items such as citizenship, health insurance, high school education, family poverty, marital status, and interruption in telephone service. In the analysis of the factual/sensitive measures, the logistic regression was constructed to predict a "sensitive" response, such as noncitizenship, lack of insurance, no high school education, and so on.

For the factual/sensitive measures, a stepped, ordered pattern was seen on citizenship, education, and poverty. In other words, as the interviewer moved into higher cooperation rate quartiles, he or she elicited more sensitive responses. However, the health insurance, marital status, and telephone service interruption variables seemed to move in the opposite direction and displayed nonordered effects.

The results became even more curious when the regression coefficients of the extended cooperation rate quartile were examined. Here, the ordered effects were evident, but moving in the opposite direction. That is, the positive impact of the coefficient was weaker for interviewers in higher extended cooperation rate quartiles (i.e., fewer sensitive responses were obtained). Again, for these variables, the coefficients for citizenship, education, and poverty were significant.

5.2 Questionnaire Items (Experience)

The analysis also examined the accumulated survey experience of interviewers in the current round of data collection. For this effort, which used contingency table analysis, cases were grouped according to the number of interviews completed by the interviewer before completing that particular case. To examine learning effects early in the experience curve, the first 20 interviews completed by the interviewer were classified into quartiles and aggregated across all cases. To examine long-term learning effects, the first 200 interviews completed by each interviewer were grouped into deciles and aggregated across all cases.

The results of this analysis showed no significant differences in the distribution of the variable groupings for the early-stage learning effect quartiles. The long-term learning effect groupings also did not display significant differences among the variable groups by interviewer classification.

5.3 Global Measures

Income Switching. In an effort to improve the precision of estimates of low-income families, the NSAF sample design included an oversample of families below 200 percent of the federal poverty level. The survey subsampled non-low-income families using a single question on the screener to determine income, and then used a series of questions on the extended interview to generate a more comprehensive estimate of family income.

Income switching occurred in the survey when a sampled unit screened in at one income level but was determined at a different income level during later in the interview. A “false negative” occurred when a household incorrectly screened in as high-income, but was determined on the extended interview to be low-income. Similarly, a “false positive” occurred when a household reported being low-income on the screener but was revealed on the extended interview to be non-low-income.

Due to the differential probabilities of selection specified by the sample design, false negatives were assigned larger weights relative to true positives (households that correctly screened in as low-income). This results in an increase in the variance of survey estimates for the low-income sample. Alternatively, false positives create sampling inefficiencies and lead to an increased cost of survey administration. For these reasons it is advantageous to minimize both the false negative and false positive rates on the survey.

In examining income-switching rates by interviewer, the regressions showed a few minor trends but no definitive patterns. False positive rates tended to increase for interviewers in higher screener cooperation rate quartiles, but paradoxically tended to decrease for the higher extended cooperation rate interviewers. Both false positive and false negative rates tended to decrease for interviewers who had worked in the first round of the survey, although these results were not significant. Experience, as measured by total number of interviews, had no discernable effect on income switching rates.

Interview Length. In computing mean interview length by interviewer, the analysis controlled for income, presence of children, number of child interviews, and presence of spouse or partner in the household, all factors that influence the number of items asked during the survey. The

regressions showed mixed results and few trends. Interviewers in higher screener cooperation rate quartiles tended to have longer average interviews; however, the opposite proved true for extended cooperation rate. While tenure appeared to be associated with conducting shorter interviews, experience seemed to have no effect.

Item Nonresponse. Finally, the analysis examined interviewer item nonresponse rates. Here again, the regressions produced few significant results, although the analysis did reveal a significant relationship between screener cooperation rates and item nonresponse on the race question. As interviewers move into higher screener cooperation rate quartiles, their item nonresponse rate on race tended to increase. Additionally, there was some indication that working in round 1 was associated with higher imputation rates across all the variables tested.

6. Discussion

6.1 Paradata

The first objective of this research was to gain knowledge on effective use of paradata, including a better understanding of the costs/benefits paradata represent to both the data producer and data user. It is clear that paradata extend a useful tool to researchers who are interested in examining the quality of data for themselves, beyond that which is communicated through response rates and sampling error.

However, the utility for the data user is tempered by the complexity of the data. Learning to use the paradata variables efficiently in order to examine potential bias of interest is a nontrivial matter and may represent a significant barrier to some researchers. Alternatively, the survey system itself may choose to shoulder the burden of providing more user-friendly summary variables based on paradata source variables, although this clearly presents an added cost to the organization, both in terms of anticipating items of interest as well as in increased programming and documentation costs. However, there is no question that as response rates continue to decline on a national level, additional measures of survey quality are of increased importance.

6.2 Interviewer Effects

After investigating the relationship between survey response and interviewer characteristics such as skill, tenure, and experience, the analysis found the patterns of response fairly similar across interviewer classifications. In the few cases in which significant differences were evident, the trends were inconsistent and inconclusive. Therefore, it was surmised that interviewer effects stemming from tenure and experience were close to undetectable using the variables employed in this analysis. However, the analysis of effects related to skill does merit further examination, particularly with regard to the observed phenomena of screener and extended interview cooperation rates having seeming opposing effects.

This analysis was limited by a number of factors. Because actual cooperation rate values are not provided, the range, or spread, both within and across cooperation rate quartiles was lost. In addition, some critical dimensions of interviewer skill were omitted. These were mostly variables that might measure the actual administration of the survey, such as percent of questions monitored not read exactly as worded, or instances of directive probing. These dimensions would

have been very attractive to analyze, but were simply unavailable in the administrative data. Finally, the restrictive controls used to build the analysis sets probably also served to reduce the variability of the estimates, but this was believed to have been a critical component of facilitating a clean look at the relationship between interviewer characteristics and characteristics of the interview.

References

Converse, Nathan, Adam Safir, and Fritz Scheuren. 2001. *1999 Public Use File Data Documentation*. Methodology Report No. 11.

Fowler, Floyd J. Jr., and Thomas W. Mangione. 1990. *Standardized Survey Interviewing*. Newbury Park, CA: Sage Publications.

Scheuren, Fritz. 2001. "Macro and Micro Paradata for Survey Assessment." In *1999 NSAF Collection of Papers*, Methodology Report No. 7.

Singer, Eleanor, M. Frankel, and M. Glassman. 1983. "The Effect of Interviewer Characteristics and Expectations on Response." *Public Opinion Quarterly* 47:68–83.

Table 1. Logit Model Results

Variable	Intercept		1st SCQ		2nd SCQ		3rd SCQ		1st ECQ		2nd ECQ		3rd ECQ		Worked C1		Race (Black)		Race (Other)		Metro Status	
	B	SE	B	SE	B	SE	B	SE	B	SE	B	SE	B	SE	B	SE	B	SE	B	SE	B	SE
<i>Subjective/non-sensitive</i>																						
Welfare Helps People	-1.39	0.23	0.18	0.16	0.21	0.15	0.27	0.12	-0.18	0.19	-0.22	0.15	0.08	0.14	-0.42	0.18	0.26	0.16	-0.52	0.33	0.00	0.13
Single Parents Effective	-0.06	0.17	-0.15	0.13	-0.07	0.11	0.00	0.10	0.07	0.11	0.02	0.10	0.00	0.08	0.05	0.13	-1.06	0.14	0.15	0.16	-0.06	0.10
Want Kids Should Marry	-2.64	0.26	-0.21	0.19	0.04	0.21	0.22	0.21	0.29	0.28	0.09	0.22	0.02	0.19	0.14	0.25	-0.34	0.27	0.35	0.33	-0.35	0.23
Working Moms Effective	-1.03	0.18	-0.07	0.14	-0.17	0.13	0.03	0.12	0.02	0.13	-0.06	0.13	0.19	0.12	-0.05	0.14	-0.44	0.14	-0.55	0.20	0.00	0.10
Welfare Works Less	-0.52	0.18	0.12	0.11	0.10	0.12	0.02	0.11	-0.30	0.17	-0.10	0.16	-0.34	0.12	-0.38	0.14	0.39	0.12	-0.06	0.23	0.02	0.10
Fair/Poor Health	-3.57	0.42	-0.33	0.27	-0.42	0.21	-0.48	0.23	0.45	0.30	0.35	0.27	0.37	0.26	0.20	0.32	0.96	0.22	0.33	0.50	0.29	0.21
Confidence in Health Care	-2.57	0.39	-0.31	0.24	-0.26	0.19	-0.21	0.19	0.60	0.34	0.20	0.25	0.04	0.25	-0.06	0.30	0.50	0.20	0.35	0.32	-0.19	0.17
<i>Subject/sensitive</i>																						
Parental Aggravation	1.92	0.31	0.13	0.17	0.12	0.19	0.10	0.18	0.22	0.27	0.21	0.18	0.09	0.17	0.29	0.26	-0.83	0.19	-0.47	0.28	0.09	0.16
Behavioral Problems (A)	2.18	0.53	0.47	0.30	0.14	0.29	-0.03	0.30	-0.44	0.38	-0.31	0.29	-0.46	0.32	0.73	0.49	-0.24	0.38	-0.68	0.71	0.18	0.27
Behavioral Problems (B)	3.24	0.49	0.35	0.37	0.71	0.32	0.40	0.30	-0.58	0.48	-0.75	0.44	-0.68	0.37	-0.48	0.38	-0.38	0.31	0.53	0.53	-0.24	0.26
School Engagement	1.12	0.22	0.60	0.16	0.32	0.16	0.44	0.14	-0.39	0.17	-0.23	0.15	-0.23	0.12	0.34	0.19	-0.29	0.16	0.20	0.36	-0.05	0.11
Negative Mental Health	1.94	0.27	0.40	0.14	0.18	0.15	0.23	0.13	-0.38	0.18	-0.34	0.13	-0.45	0.11	0.10	0.21	-0.36	0.14	-0.13	0.29	-0.28	0.14
<i>Factual/sensitive</i>																						
Citizenship	-2.96	0.35	-1.53	0.31	-1.51	0.20	-0.66	0.17	1.24	0.26	1.06	0.23	0.90	0.25	0.39	0.26	-0.59	0.21	0.79	0.27	-1.06	0.29
Health Insurance	-2.36	0.37	-0.25	0.20	-0.34	0.16	-0.33	0.14	0.36	0.22	0.18	0.17	0.29	0.18	0.06	0.31	0.23	0.17	-0.32	0.30	0.21	0.15
HS Education	-3.10	0.28	-0.64	0.23	-0.60	0.20	-0.42	0.16	1.09	0.22	0.93	0.17	0.94	0.17	0.46	0.24	0.24	0.17	0.07	0.27	0.17	0.16
Poverty	-0.87	0.18	-0.44	0.12	-0.27	0.13	-0.27	0.10	0.58	0.15	0.27	0.13	0.24	0.12	-0.04	0.15	1.09	0.11	-0.12	0.20	0.54	0.08
Marital Status	-1.49	0.21	-0.03	0.12	-0.03	0.10	-0.18	0.11	0.16	0.15	0.14	0.11	0.03	0.11	0.27	0.17	1.70	0.11	0.24	0.24	-0.01	0.10
Telephone Interruption	-2.94	0.35	-0.08	0.23	-0.16	0.21	-0.17	0.21	0.32	0.18	0.10	0.20	0.25	0.16	-0.03	0.29	1.33	0.16	0.22	0.36	0.65	0.18
Child Working	0.65	0.32	0.11	0.19	-0.15	0.19	0.26	0.19	0.07	0.24	0.15	0.19	0.09	0.19	-0.01	0.27	1.22	0.27	1.09	0.42	-0.17	0.14
Suspended/expelled	-2.09	0.41	0.05	0.31	-0.37	0.33	-0.03	0.21	0.17	0.34	0.14	0.27	0.44	0.30	-0.25	0.32	0.69	0.27	0.66	0.47	0.26	0.19

Notes: Figures in bold are statistically significant at $p < .05$. Significance not shown for control variables (race and metro status).

Table 2. OLS Model Results

Variable	Intercept		1st SCQ		2nd SCQ		3rd SCQ		1st ECQ		2nd ECQ		3rd ECQ		Worked C1		Total Int	
	B	SE	B	SE	B	SE	B	SE	B	SE	B	SE	B	SE	B	SE	B	SE
<i>Income Switching</i>																		
Total False Negative Rate	12.54	1.26	-1.06	1.42	-1.40	1.45	0.49	1.40	0.50	1.53	0.04	1.38	-0.08	1.39	-0.36	1.49	-0.01	0.01
Total False Positive Rate	28.74	1.67	1.02	1.91	1.47	1.93	1.93	1.87	1.20	2.04	-0.13	1.86	-0.34	1.86	-1.25	1.99	0.00	0.01
False Negative Rate Same Resp	11.16	1.34	-1.64	1.51	-2.53	1.55	-0.46	1.49	1.41	1.62	-0.92	1.47	-0.39	1.47	0.31	1.59	-0.01	0.01
False Positive Rate Same Resp.	25.44	1.76	1.33	1.99	0.51	2.02	2.16	1.95	-0.31	2.14	-0.26	1.95	-1.05	1.94	-1.55	2.07	0.01	0.01
False Negative Rate Diff. Resp.	15.30	2.13	-1.42	2.37	0.24	2.43	1.59	2.33	0.56	2.54	3.67	2.31	1.60	2.31	-0.15	2.47	0.00	0.01
False Positive Rate Diff. Resp.	34.25	3.04	-1.29	3.32	2.30	3.41	1.50	3.26	7.02	3.60	3.88	3.27	2.68	3.27	-2.70	3.42	-0.01	0.02
<i>Average Interview Length</i>																		
Avg. Int. Length High Income	41.57	0.82	-1.83	0.89	-2.51	0.92	-1.66	0.88	-0.10	0.96	1.33	0.88	0.48	0.88	-0.30	0.92	-0.01	0.00
Avg. Int. Length Low Income	48.65	1.55	-3.11	1.67	-2.70	1.71	-2.20	1.66	2.99	1.84	2.99	1.67	1.90	1.68	-3.47	1.73	-0.02	0.01
<i>Item Non-Response Rates</i>																		
UBRACE4	9.26	1.47	-5.60	1.67	-5.08	1.70	-3.97	1.64	1.52	1.79	1.94	1.62	1.26	1.62	0.62	1.76	0.00	0.01
IHRAMT	6.97	0.64	0.77	0.73	0.95	0.75	0.08	0.72	-0.58	0.79	-0.70	0.71	-0.96	0.71	0.41	0.77	0.00	0.00
IPAYAMT	6.69	0.65	0.88	0.74	0.52	0.76	0.77	0.73	-0.50	0.80	-0.12	0.72	0.40	0.72	0.84	0.78	0.00	0.00
ISETOT	2.62	0.38	-0.42	0.43	-0.29	0.44	-0.23	0.43	0.45	0.47	0.36	0.42	0.12	0.42	0.34	0.46	0.00	0.00

Using a Short Follow-up Survey to Compare Respondents and Nonrespondents

Timothy Triplett, Adam Safir, Kevin Wang, Rebecca Steinbach, and Simon Pratt

1. Introduction

This paper presents an analysis of the potential for nonresponse bias in the 1999 National Survey of America's Families (NSAF), a survey of the well-being of children, adults under the age of 65, and their families. The NSAF is primarily a random digit dial (RDD) telephone survey, consisting of a short screener interview to determine household eligibility and a longer extended interview during which survey items of interest are gathered for sampled household members. In order to examine the potential for nonresponse bias, a follow-up survey of a sample of respondents and nonrespondents from the NSAF screener interview was conducted by a different survey organization than the one that conducted the main survey. To measure differences between respondents and nonrespondents, the follow-up survey included key items from the main survey. In addition, the follow-up survey contained questions on subjects that were thought to be correlated with willingness to participate in a survey, such as attitudes toward surveys and government, and perceptions of being busy.

2. NSAF Survey—"The Main Survey"

The NSAF survey is funded by a consortium of private foundations in the United States and is conducted by Westat for the Urban Institute. The purpose of the survey is to assess the impact of recent changes in the administration of a number of assistance programs for children and the poor.

The sample is based on two different frames. The largest portion of the sample is RDD and is used to represent households with telephones. An area frame is used to select households that do not have telephones. All interviews were done by telephone, with interviews in the area frame being conducted using cellular phones supplied to the respondent. The sample of the NSAF is designed to generalize to 13 specific states, as well as the country as a whole. There is also an oversampling of households that were estimated to be under 200 percent of the federal poverty level.

The NSAF consists of both a screening and an extended interview. The screening interview consists of about three minutes of questions designed to assess eligibility and select the person that should be administered the extended interview. This involves determining whether there are any persons under 65 years old in the household and whether or not the family is above or below 200 percent of the poverty level. If there is someone in the right age-range and the household is sampled (based on poverty status), a respondent for the extended interview is selected. The extended interview is between 25 and 45 minutes in length and covers a wide range of topics, including health, education, child care, income, and receipt of social services.

The response rate for the screener interview was 76.7 percent. The final combined response rate (screener response rate multiplied by the extended response rate) ranged from 61 to 67 percent,

depending on the type of interview (adult versus family). A total of 46,705 extended interviews were completed between February and October 1999.

3. University of Maryland—“The Nonresponse Follow-up Survey”

The data collection period for the follow-up survey was between August 25, 1999, and October 18, 1999. Therefore, the follow-up study took place while the NSAF survey was still being completed.

The sample for the follow-up survey consisted of 2,000 finalized telephone numbers from the 1999 NSAF study. The selection of telephone numbers for the follow-up survey was done using an equal probability sample within the following three NSAF screener outcome groups: households that completed the screener without ever refusing ($n = 500$); households that completed the screener survey but initially refused ($n = 600$); and households that were finalized as nonrespondents to the NSAF screener ($n = 900$). Nonrespondents were mostly refusals, but also included those who received the maximum number of calls according to study protocol, those who never answered but had answering machines, and other nonresponse in the NSAF main study. Some telephone numbers were excluded from the experiment: language problem cases, nonworking cases, nonresidential business cases, nonworking tritone matches (determined by a computer system that dialed telephone numbers to detect the tritone signal and eliminate those that were nonworking), duplicate cases, hostile refusals, and cases for which the telephone was never answered (NA).

The follow-up questionnaire included some key NSAF questions and demographics questions that were also asked on the NSAF questionnaire (food stamps, health insurance, household composition, education, employment, race, ethnicity). Other questions were added to obtain information that could explain nonresponse, such as respondents' opinions about the importance of research surveys, how much time they feel they have, and how they feel about opinion pollsters. Follow-up respondents were also asked about their opinions about government. In addition, questions were asked about the number of telephone numbers in the household that were used for non-voice communication (e.g., telephone lines used only for computers). The average time to complete the interview was 8.3 minutes.

Data collection for the follow-up study was conducted by the University of Maryland's Survey Research Center. A major reason for having a different data collection organization for the follow-up study was the need to have the follow-up study be as independent of any issues that may have arisen in the original NSAF survey as possible. Sample cases in the main study were mailed letters, brochures, and incentives, and were called repeatedly to obtain their cooperation. By having a different data collection organization perform the follow-up, it was hoped that some of the effects of these efforts could be isolated. To further this objective, the sponsor of the follow-up survey was also changed. Child Trends was the sponsor for the follow-up survey.

Respondent selection was the same in both the main and follow-up survey. Any adult member of the household could complete the NSAF screener and likewise the follow-up survey could be completed by any household member 18 years of age or older. Some of the other important features for the follow-up study are described below.

Number of calls. Cases were called nine different times on different days.

Spanish language. Bilingual interviewers were hired and trained for Spanish-speaking households.

Refusal conversion. Refusals were held for 10 days, and one refusal conversion attempt was made.

Letters. Letters were sent to aid in refusal conversion. In order to distinguish this letter from those associated with the main NSAF study, money was not included and the letters were sent priority mail rather than by an overnight service, as they were in the main study. The letters were on the University of Maryland's letterhead.

4. Response Rate Differences

Table 1 shows three different response rate calculations. The first column is what is usually described as the cooperation rate, which is simply the total completes for the follow-up survey divided by the total completes plus refusals. The second column is the response rate, which is the total completes divided by total completes plus the refusals and the other nonrespondent eligible households (home recorders, max calls on callbacks, language and health problems). The third column is labeled the completion rate; this is the percentage of all phone numbers provided for which a completed interview was obtained (completes divided total sample provided).

Table 1. Survey Rates by Strata

	<i>N</i>	Cooperation rate	Response rate	Completion rate
Completed and never refused	600	85.3%	78.1%	70.2%
Completed and initially refused	500	63.1%	59.7%	51.0%
Nonrespondents	900	37.9%	29.7%	25.7%
Nonrespondent refusals	754	35.3%	30.0%	26.4%
Nonrespondent Max Calls	146	69.6%	28.0%	21.9%

For the most part, the cooperation, response and, completion rates are what one would expect to occur. The respondents who completed the screener without ever refusing were the most likely group to cooperate and respond to the follow-up survey. Respondents that initially refused the NSAF screener were slightly less cooperative, but much more cooperative than the nonrespondents who refused the NSAF. The nonrespondents that did not refuse the NSAF screener were actually more cooperative than the those who completed the NSAF after initially refusing. However, this max call and other nonrespondent group was actually the most difficult group to complete follow-up surveys with (21.9 percent), even though they were unlikely to refuse the survey. This could be an indication that RDD telephone surveys are sometimes misclassifying telephone numbers as nonresponse, when in fact the number may not be associated with a residential household.

5. Adjustments Made Prior to Analysis

First, we decided not to include the 32 follow-up interviews that were completed with nonrespondents that did not refuse the NSAF screener. In addition to the differences in

cooperation, other research has shown that this group differs from that of respondents who refuse (Black and Safir 2000; Triplett 2001; Groves and Couper 1998). Therefore the nonresponse group ($n = 209$) consists of only respondents who completed the follow-up survey but refused the NSAF.

Second, our analysis used a weighting factor that controlled for both the differential sampling within stratum and a follow-up study nonresponse adjustment. Thus, after applying the weight, the percentage of respondents who completed the follow-up study were proportional to the percentage of respondents who either completed the NSAF screener, completed the screener after refusing, or refused and never completed.

Third, we decided to collapse the two groups that completed the NSAF and use completing or refusing the NSAF as the dependant variable in our logistical regression analysis. This was done since our primary objective was estimating the potential for nonresponse bias in the NSAF. In addition, we also found that respondents who completed the NSAF but initially refused were more like the initial cooperators than the nonrespondents. This finding is supported by the research on the impact of nonparticipation done by Lin and Shaeffer (1995).

6. Analysis of the Behavioral and Attitudinal Questions

In designing the follow-up questionnaire, a number of behavioral and attitudinal questions were asked, thinking that they would help explain nonresponse. In total there were 10 of these questions asked during the follow-up survey (table 2). The order in which questions 2, 3, and 4, and questions 6 through 10 were asked was rotated to reduce the effect of any bias that may occur due to the order in which the questions were read.

Table 2. Behavioral and Attitudinal Questions

1. How important do you think it is that research is done about education, health care, and services in your [fill STATE]. Would you say it is:	
Extremely important,	1
Very important,	2
Somewhat important,	3
Not too important, or.....	4
Not important at all?	5
For each of the following statements, please tell me if you strongly agree, somewhat agree, somewhat disagree, or strongly disagree with each statement.	
2. Research surveys help improve the way government works.	
3. People like me don't have any say about what the government does.	
4. People have lost all control over how personal information about them is used.	
5. In general, how do you feel about your time? Would you say you always feel rushed even to do things you have to do, only sometimes feel rushed, or almost never feel rushed?	
Now I'd like you opinion of some people and organizations. As I read from a list, please tell me which category best describes your overall opinion of who or what I name.	
6. Would you describe your opinion of Congress as...	
7. Would you describe your opinion of the Democratic party as...	
8. Would you describe your opinion of the Republican party as...	
9. Would you describe your opinion of pollsters as ...	
10. Would you describe your opinion of telemarketers as ...	
Very favorable,	1
Mostly favorable,	2
Mostly unfavorable, or.....	3
Very unfavorable?	4

In comparing the mean scores for each of these 10 questions (table 3) by the three NSAF response and nonresponse groupings (initial cooperators, initial refusals and nonrespondent refusals), only the question that asks respondents about their opinion of pollsters provided statistically significant findings. This was somewhat surprising, since most of these items were expected to produce some differences between respondents, reluctant respondents, and final NSAF refusals.

The NSAF refusal group had a significantly lower opinion of pollsters than both the reluctant respondents and respondents. In addition, while all groups usually gave very unfavorable ratings to telemarketers, the refusal group was the least unfavorable. Thus, it appears that those who refused to participate in the NSAF study but did the follow-up study do not make as much of a distinction between telemarketers and pollsters. This finding could be very problematic for surveys that either define themselves or are perceived as opinion polls. University and other research-sponsored studies may want to avoid using terms that make them sound like they are conducting an opinion poll.

Since all these behavioral and attitudinal questions were items that were thought to be predictors of nonresponse, there was correlation found among the questions. Therefore, to further test the significance of the opinion of pollsters, we decided to run logistic regression using the behavioral

and attitudinal questions as predictors of nonresponse. The results of this regression analysis (first column of table 4, Model 1) support the finding that nonrespondents did have a significantly different opinion of pollsters, while the difference on the other items remained insignificant. While not quite significant at the .05 level, the regression also supports the finding that less cooperative respondents have a higher opinion of telemarketers relative to pollsters.

7. Comparison of Demographic Characteristics

We did find some differences when comparing the demographic characteristics of the NSAF refusals with those who completed NSAF. Those who completed the NSAF were more likely to be white and own their own home, but were also more likely to be unemployed. Those who refused were more likely to be from larger households and black, but were also more likely to have graduated high school or received their GED.

When we used the demographic variables in our logistic regression analysis (table 4, Model 2) to try to predict whether a person completes the NSAF, the employment variable was no longer found to be significant. However, we did find that adults in the household, homeownership, and high school degree or GED were significant predictors of responding to the NSAF. Race was significant at the .1 level, but not at the .05 level.

Do the differences we found in the respondents and nonrespondents demographic characteristics help explain this difference we found in their opinion of pollsters? In order to answer this question, we combined in our logistical regression analysis (table 4, Model 3) the demographics variables that were found to be significant predictors of responding to NSAF response with the respondent's opinion of pollsters and telemarketers. This combination in fact slightly increased both while reducing the estimates standard error, thus further strengthening the argument that respondents and nonrespondents differ in their opinion of pollsters.

Table 3 Mean Scores

	Initial cooperators (n = 426)	Initial refusals (n = 260)	Refusals (n = 207)
Importance of research	1.77	1.81	1.91
Research surveys help	2.14	2.23	2.23
Cannot change government	2.64	2.53	2.52
Lose control of personal info	1.88	1.94	1.89
Feel rushed	1.83	1.87	1.74
Opinion of Congress	2.42	2.47	2.51
Opinion of Democrats	2.43	2.37	2.56
Opinion of Republicans	2.58	2.53	2.6
Opinion of pollsters	2.41	2.54	2.69
Opinion of telemarketers	3.62	3.62	3.51

Table 4. Logistical Regression

	Model 1		Model 2		Model 3		
	Beta	SE	Beta	SE	Beta	SE	
Intercept	-0.222	0.755	2.334	0.43	0.962	0.544	
Importance of research	-0.043	0.111					
Research surveys help	-0.102	0.127					
Cannot change government	0.031	0.083					
Lose control of personal info	0.006	0.107					
Feel rushed	0.213	0.137					
Opinion of Congress	-0.063	0.149					
Opinion of Democrats	0.078	0.119					
Opinion of Republicans	0.138	0.127					
Opinion of pollsters	0.328	0.133			0.334	0.111	
Opinion of telemarketers	-0.040	0.130			0.045	0.124	
Employed			-0.197	0.189			
Hispanic			0.192	0.307			
HS/GED			-1.065	0.387	*	-0.851	0.377
Spouse present			-0.091	0.187			
Black			-0.467	0.278		-0.430	0.281
Homeowner			0.604	0.199	*	0.698	0.202
Number of adults			-0.235	0.078	*	-0.239	0.079
Kids 0–5			0.012	0.118			
Kids 6–17			0.015	0.090			
Foreign-born			-0.040	0.312			

Notes: Dependent variable: 0 = refused NSAF, 1 = completed NSAF

* Difference significant at the .05 level.

8. Analysis of Key NSAF Questions

The follow-up survey included several items that are important to researchers who use the NSAF data. For instance, whether a family is above or below 200 percent of the poverty level; does anyone in the household receive food stamps; does the respondent or any children in the household not have health insurance. We found virtually no differences on these items when comparing the NSAF respondents and nonrespondents. This finding reduces the potential impact of nonresponse bias for much on the NSAF analysis.

9. Summary

In order to improve future surveys, the industry needs to address the apparent paradox that while respondents think positively of the contribution surveys make to improving government, they think negatively of the people who collect the data. Both respondents and nonrespondents generally favored the use of surveys, but both groups had a low opinion of pollsters and telemarketers. The opinion of pollsters was significantly worse among the NSAF nonrespondents, suggesting that some people still respond to telephone surveys and do not feel quite as negatively toward pollsters. Thus, response rates would increase if we could improve the overall perception respondents have of the data collectors.

If attitudes toward pollsters are related to survey items, it is likely that post-stratification weights would not adjust for the potential bias. Post-stratification usually involves making weighting adjustments based on a person's demographic characteristics, which we found to be independent of peoples attitudes toward pollsters. If attitudes toward pollster are correlated to any questions you might ask during a telephone survey, there is likely to be some bias in your estimate that will be difficult to account for. Fortunately for the users of the NSAF data, we found no correlation between nonresponse and the key NSAF items.

The finding that nonresponse had little impact upon the NSAF key estimates is further supported by the work of Scheuren (2000). Scheuren used a capture/recapture model to score nonresponse adjustments and found that 60 percent of the NSAF screener nonresponse is ignorable. A follow-up survey of respondents and nonrespondents was also conducted for the 1997 NSAF study. In analyzing the results from the 1997 follow-up survey, Groves and Wissoker (1999) found that NSAF nonrespondents tend to be black non-Hispanics. While the post-stratification adjustments achieve census-based representation, there is potential bias if black nonrespondents are financially worse off then black respondents. While the current follow-up survey found black non-Hispanics more likely to be nonrespondents, this finding was not significant at the 95 percent confidence level.

References

- Black, Tamara, and Adam Safir. 2000. "Assessing Nonresponse Bias in the National Survey of America's Families" In *Joint Statistical Meetings 2000 Proceedings Section on Survey Methodology*.
- Brick, Mike, Nancy Vaden-Kiernan, Carl Fowlkes, and Pat Cunningham. 2001. "Report on the NSAF Cycle 2 Capture/Recapture Study." In *1999 NSAF Collection of Papers*, Methodology Report No. 7.
- Groves, Robert, and Mick Couper. 1998. *Nonresponse in Household Surveys*. New York: John Wiley.
- Groves, Robert, and Douglas Wissoker. 1999. *Early Nonresponse Studies of the 1997 National Survey of America's Families*. Methodology Report No. 7.
- Len, I-Fen, and Nora Shaeffer. 1995. "Using Survey Participants to Estimate the Impact of Nonparticipation." *Public Opinion Quarterly* 59:236–58.
- Scheuren, Fritz. 2000. "Quality Assessment of Quality Assessment." In *American Statistical Association 2000 Proceedings of the Section on Government Statistics and Section on Social Statistics* (44–51). Alexandria, VA: American Statistical Association.
- Triplett, Timothy. September 2001. "What Is Gained from Additional Call Attempts and Refusal Conversion, and What Are the Cost Implications?" Unpublished ongoing research paper. <http://mywebpages.comcast.net/ttriplett13/tncpap.pdf>.

Sampling Refusals: Why, When, and How Much?

Timothy Triplett, Adam Safir, Kevin Wang, and Natalie Abi-Habib

1. Introduction

The 2002 National Survey of America's Families (NSAF) is a dual frame survey that relies primarily on a large RDD sample design using over 500,000 telephone numbers. The survey consists of a short three-minute screener survey used to determine eligibility followed by a 45-minute extended interview. Although interviews are completed with close to 40 percent of all initial refusals, the per-interview cost of converted refusals far exceeds that of initial cooperators. In addition, the data collection period is lengthened by refusal conversion. Almost half of all potential respondents initially refused to participate, creating cost and scheduling problems. These problems contributed to the decision not to attempt to convert screener refusals in the final 20 percent of the NSAF 2002 sample release. This strategy for dealing with nonresponse is sometimes referred to as "double sampling" or "two-phase sampling." The purpose of using a double sampling approach is that it offers a method of balancing costs and errors in deciding what efforts should be made to measure sample persons who refuse initial interview attempts (Groves 1989). While we will mention some of the literature that has tried to quantify the balancing of costs and errors, this paper will focus on possible auxiliary effects to the cooperation, response, and refusal conversion rates as a result of the double sampling approach.

2. Double Sampling

The double sampling theory has been around for a long time (Neyman 1938; Deming 1953; Hansen and Hurwitz 1946). A double sampling strategy is appealing when the total cost of a survey must be within the amount appropriated, and when some clearly defined portion of your population of interest is more difficult to interview. The double sampling strategy as it pertains to this research involves dividing the sample into two groups, those who initially refuse to do the survey and those who do not refuse. The cost and time associated with completing interviews in households that initially refuse is assumed to be greater. You would attempt to complete an interview in 100 percent of the households that never refused, while attempting to complete an interview in some fraction (80 percent in the NSAF) of households that initially refused. Resources saved from having to do less refusal conversion will effectively increase the overall size of the sample that can be worked given a fixed budget. This increased sample size will improve the precision of the survey estimate. However, for the final sample to be representative, the initial refusal sample should be weighted by the inverse of the sampling fraction (1.25 in the 80 percent example). This additional weighting factor will increase the variance of the survey estimate. Choosing the optimal fraction of refusals to call back depends on the trade-off between increasing final sample size and increasing the variance associated with the weighting adjustment from sampling refusals. The difficulty with using a double sampling approach for dealing with refusals is that choosing an optimal sampling fraction depends on cost and error information that is not known before data collection. Further complicating matters is that some other factors could be associated with the double sampling process itself that could alter cost and error estimates. Deciding not to call refusals may affect interviewer morale, interviewer's perception of the

importance of not getting a refusal, and could also affect interviewer work assignments. There are probably other factors that may affect the cost and error estimates from deciding to sample households that initially refused. The main purpose of this paper is to see how much of an impact other factors may have had on the double sampling process carried out in the 2002 NSAF.

3. NSAF Sample Design

The purpose of the NSAF is to assess the impact of recent changes in the administration of a number of assistance programs for children and the poor. The NSAF sample is designed to generalize to 13 specific states, as well as to the nation as a whole. The design also includes an oversampling of households that were estimated to be under 200 percent of the federal poverty level as well as households with children. All three rounds of NSAF data collection (1997, 1999, and 2002) were done for the Urban Institute by Westat.⁸

The NSAF consists of both a screening and an extended interview. The screener consists of about 3 minutes of questions designed to assess household eligibility and select a respondent for the extended interview. The sampling of initial refusals applied only to the screening interview. The RDD telephone sample for the NSAF was randomly divided into 101 replicates. The first 81 replicates were designated as the refusal conversion sample while the final 20 replicates were designated as the nonrefusal conversion sample. Therefore, the fraction of nonrespondents who received standard refusal conversion was actually slightly higher than 80 percent (81/101 replicates). The no-refusal conversion sample (last 20 replicates) was released by August 2002 and data collection ended on November 3, 2002, therefore all telephone numbers in the 101 replicates were able to receive their full complement of call attempts.

4. Possible Auxiliary Effects

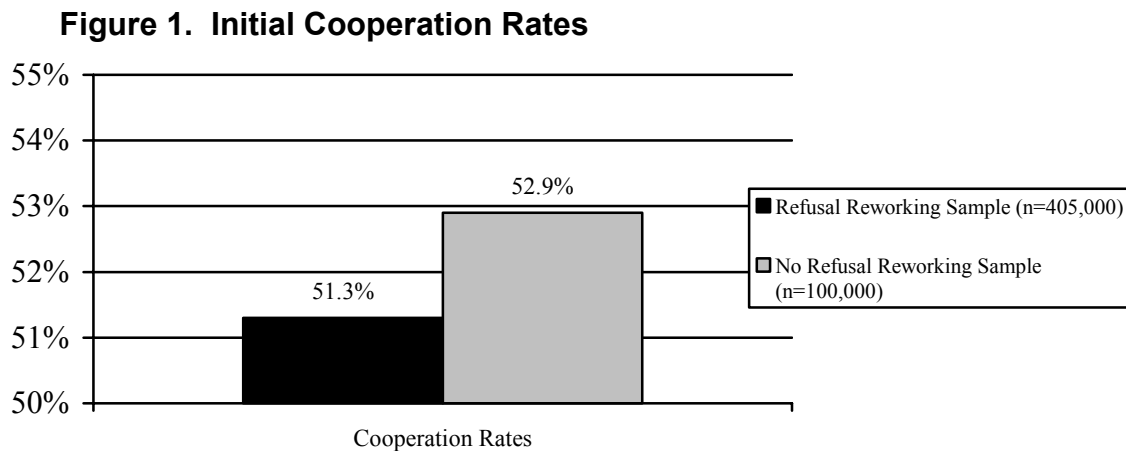
If sampling refusals did affect the survey process, you would expect to see differences in the initial cooperation rate between the first 81 replicates where standard refusal conversion efforts were carried out versus the final 20 replicates where refusal conversion was not done. Additionally, you might expect a change to occur in the refusal conversion rate, since fewer refusals will be called back, which would affect the interviewers work assignments.

We used the NSAF data to test whether sampling nonrespondents affects the initial cooperation or refusal conversion rates. There are two cooperation rate comparisons of interest. First, there is the simple comparison of initial cooperation rates from the refusal conversion sample versus the nonrefusal conversion sample. Second, we tested for an effect on the initial cooperation rate of the refusal conversion sample in correlation with the release of the no refusal conversion sample. Testing for an effect on refusal conversion rates involves comparing refusal conversion rates before and after the release of the nonrefusal conversion sample.

⁸ The screener response rate has declined from 77% in 1997 to 66% in 2002.

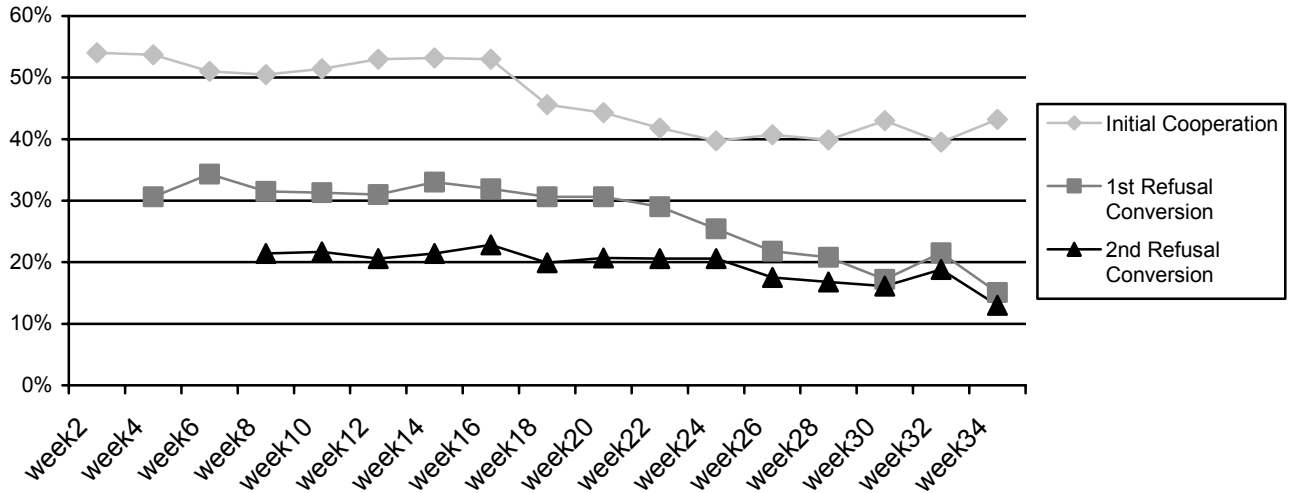
5. Results of the Tests

The first test compared initial cooperation rates in both samples, first the 81 replicate releases (405,000 telephone numbers) in which refusals were reworked versus the 20 replicates (100,000 telephone numbers) in which refusals were not called. As seen in figure 1, the initial cooperation rate is higher for the nonrefusal conversion sample. The increase from 51.3 percent to 52.9 percent is statistically significant, although not large. This increase would support the hypothesis that sampling nonrespondents could improve morale or interviewer staffing. However, given how small the increase is overall, it could also be explained by factors unrelated to the double sample process, such as changes in interviewers skills and morale that evolve over time.



Determining whether the initial cooperation rate in the convert refusal sample was affected by the introduction of the sample where no refusals were reworked is more difficult. This is because the initial cooperation rate of RDD sample usually begins to decline as the number of call attempts needed to reach a household increases (Lynn et al. 2002; Triplett 2002). However, there is some evidence that the initial cooperation rate in the convert refusal sample dropped more dramatically than expected. This can be seen by the sudden drop in the initial cooperation rate indicated by the arrow in figure 2. At week 18 of NSAF data collection, the no-refusal conversion sample began being called by the interviewers. Therefore, any effect from the change in how the sample is worked would be expected to occur shortly thereafter. It is generally expected that there would be a decline in the initial cooperation rate of older sample any time an additional sample is released. However, the decline in week 18 was larger than occurred during any other week of the study, and there were several other times during the study when a new sample was released. Thus, it appears that giving interviews a fresh sample that was not going to require refusal conversion had unintended negative effects on the initial cooperation rate of the older sample. Was this decline a result of reduced enthusiasm for making call attempts on the older sample that requires refusal reworking? While this could be an explanation, it is not something we are able to test for.

Figure 2. Refusal Reworking Time Line



Testing if the refusal conversion rate benefited from the reduction in total refusals needing rework is also quite difficult to measure. The effects from having fewer refusals to convert would likely occur a few weeks after the no-refusal conversion sample was released, since conversion attempts require at least a two-week waiting time on the NSAF. We would expect softer refusals would be converted sooner, thus over time, the refusal conversion rate would fall. The refusal conversion rate for both first and second refusals gradually decreases over time (figure 2). Since the no-refusal conversion sample was released in week 18, the benefit from having fewer refusals to rework would have begun by week 20. It appears that having fewer refusals to convert that may have helped refusal conversion. In week 20, the refusal conversion rate on first refusals stopped declining, while the conversion rate for second refusals increased slightly. These changes are not large and could have other explanations, but they support the hypothesis that by reducing the number of refusals that need converting, it becomes easier to assign the better interviewers to rework refusals.

6. Summary Discussion

We have explored some of the subtler effects of sampling nonrespondents. A more important issue in deciding whether to sample refusals and how many is how well refusal conversions improve estimates. Recent studies have found little or no reduction in nonresponse bias from efforts at increasing response rate (Lynn et al. 2002; Teitler et al. 2003; Curtin et al. 2000; Keeter et al. 2000). These findings suggest that the fraction of refusals that should be called back could be set much higher than the 20 percent chosen for the NSAF without affecting survey estimates.

One thing we know from the NSAF data collection is that we were able to finish the study earlier than we would have had we not sampled refusals. We also were able to increase the amount of overall sample released. It is less clear whether we benefited from higher initial cooperation rates due to sampling refusals, but we certainly did not do any worse. Since we achieved a higher cooperation rates from the no-refusal conversion sample, the weight adjustment (total initial

refusals/initial refusals from the sample receiving conversion attempts) was slightly smaller than expected. A smaller weighting adjustment reduces the variance associated with this weight.

While our findings on the effects of sampling refusals on cooperation and refusal conversion rates were not very strong, this could be in part due to this not having been a controlled experiment. A controlled experiment with careful interviewer assignments would have improved our ability to test for effects on cooperation rates from sampling refusals. While a controlled experiment might provide additional evidence of auxiliary effects from sampling refusals, it is still likely that these effects will be small and even less likely that they would wind up negatively affecting the survey. With respondents increasingly refusing surveys, it is likely that there will be more double sampling as way of handling nonresponse in the future. Therefore, more research is needed to assess the effects of sampling nonrespondents on the overall survey error.

References

- Brooks, Camilla A., and William D. Kalsbeek. 1982. "The Double Sampling Scheme and Its E(MSE)." In *Proceedings of the Section on Survey Research Methods* (235–39). Alexandria, VA: American Statistical Association.
- Curtin, Richard, Stanley Presser, and Eleanor Singer. 2000. "The Effect of Response Rate on the Index of Consumer Sentiment." *Public Opinion Quarterly* 64(4): 413–28.
- Deming, Edwards W. 1953. "On a Probability Mechanism to Attain an Economic Balance between the Resultant Error of Response and the Bias of Nonresponse." *Journal of the American Statistical Association* 48(264): 743–72.
- Elliott, M.R., and Roderick J.A. Little. 2000. "Subsampling Callbacks to Improve Survey Efficiency." *Journal of the American Statistical Association* 95:730–38.
- Groves, Robert M. 1989. *Survey Costs and Survey Errors*. New York: Wiley.
- Hansen, Morris H., Hurwitz. 1946. "The Problem of Non-Response in Sample Surveys." *Journal of the American Statistical Association* 41(236): 517–29.
- Keeter, Scott, Carolyn Miller, Andrew Kohut, Robert M. Groves, and Stanley Presser. 2000. "Consequences of Reducing Nonresponse in a National Telephone Survey." *Public Opinion Quarterly* 67(2): 125–48.
- Lynn, Peter, Paul Clarke, Martin, and Paul Sturgis. 2002. "The Effects of Extended Interviewer Efforts on Nonresponse Bias." In *Survey Nonresponse*, edited by Robert M. Groves, Don Dillman, John L. Eltinge, and Roderick J.A. Little (135–47). New York: Wiley.
- Neyman, J. 1938. "Contribution to the Theory of Sampling Human Populations." *Journal of the American Statistical Association* 33(201): 101–16.
- Teitler, Julien O., Nancy E. Reichman, and Susan Sprachman. 2003. "Costs and Benefits of Improving Response Rates." *Public Opinion Quarterly* 67(1): 126–38.
- Triplett, Timothy. September 2001. "What Is Gained from Additional Call Attempts and Refusal Conversion, and What Are the Cost Implications?" Unpublished ongoing research paper. <http://mywebpages.comcast.net/ttriplett13/tncpap.pdf>.

Trimming Extreme Weights in Household Surveys

Benmei Liu, David Ferraro, Erin Wilson, and J. Michael Brick

1. Introduction

An outlier in sample surveys occurs when an extreme value is observed for a characteristic, the weight for a sampled unit is very large relative to the weights of other units, or there is a combination of a relatively unusual observed value and a large weight. To be considered an outlier, the case should be influential—in the sense that it has a major effect on data analysis. However, the influence might be in estimating the variance of the estimate and, thus, affecting inferences. If an observed value is unusual and the sampled unit has a relatively small weight, then the influence of the case on most analyses may not be sufficient to classify it as an outlier.

Outliers may appear even if the sample design, data collection, and data preparation are carefully crafted and implemented. As noted by Lee (1995), outliers in sample surveys may be either representative or nonrepresentative. Representative outliers are correctly recorded and represent other population units similar in value to the observed outliers. Nonrepresentative outliers are those that are incorrectly recorded or unique in the sense that there are no other units like them. We consider only representative outliers. In addition, it is worth noting that observations that are not outliers when full population estimates are produced may be outliers for estimates of domains. This further complicates the problem of identifying outliers.

This paper focuses on methods of identifying cases in household surveys as outliers because they have large weights, rather than dealing with unusual observed values of characteristics. There are several reasons for this focus. First, large weights are most likely to have a substantial influence for a variety of analyses, especially for estimates of domains. Second, at the time of data processing, it is possible to identify large weights and trim or otherwise deal with them. In some circumstances this may not be possible with the observed characteristics. Third, many existing procedures recommend ways of identifying and dealing with observed values that are unusual. These procedures generally do not deal with unequal weights. Thus, there is a greater need to deal with the identification of influential weights in a sampling setting.

Outliers due to extremely large weights most often occur when the design samples units with different probabilities to meet some design goal. For example, a survey may sample household members with different probabilities to achieve specified sample size goals by age, or adults may be sampled with different probabilities across households so no more than one adult is selected within a household. If outliers can be identified during the weighting stage of the survey, it is possible to make an adjustment, such as trimming the weight, that will eliminate some problems during subsequent analysis of the data.

Potter (1988, 1990) and Lee (1995) review and propose methods for dealing with outliers in sample surveys, where the outliers may be due to unusual observed values or large sampling weights. Despite the suggestions they and others have given, most surveys still use ad hoc procedures that may be arbitrary and not supported by any theory in order to identify outliers. One reason for this may be that the methods suggested in the literature are not fully defined, in

the sense that they do not provide specific criteria for classifying observations as outliers. Consequently, the evaluation of alternative methods is very limited.

The next section examines methods of identifying outliers in household sample surveys. We begin by reviewing some methods suggested in the literature and discussing some of their limitations. New methods are then proposed that are more relevant to identifying cases that have large and influential weights. The methods rely heavily on the ranking of the weights. We propose specific procedures to identify outliers that may be portable across different household survey designs and may eliminate some of the arbitrariness often associated with this task. Thus, guidelines for classifying units with large weights as outliers are proposed. The third section implements the proposed methods using two complex surveys conducted by Westat. The strengths and the weaknesses of the methods are evaluated. The last section gives some conclusions and suggestions for additional research.

2. Outlier Detection Methods

2.1 Existing Methods from Literature Review

The literature for identifying and modifying outliers in household survey data is relatively limited. Most methods have been designed to handle survey specific situations and cannot be successfully applied to general household survey designs. Some methods are discussed in this section, as well as new methods devised throughout the course of this research. Some methods could be used to identify very large weights *and* very small weights. As discussed in the introduction, only large weights are considered in this study because small weights do not have as much of an effect on sample estimates.

An extreme weight may be declared as an outlier based on its relative distance from the center of the data. For instance, let $distance_i = \frac{|w_i - m|}{s}$, where w_i is the weight for sampled unit i , m is a location measure representing the center of the weights, such as the median, and s is a scale measure. One common candidate for scale measure is the median absolute deviation, defined as

$$AD = median \left\{ \left| w_i - median_j(w_j) \right| \right\}.$$

When an observation has a large value of $distance_i$, it indicates that the weight is relatively large compared with the other weights in the dataset.

Another method for identifying outliers is the forward search method described by Chambers (2003). Assuming a survey data set of size n , the algorithm begins with an initial subset of $m < n$ observations that are assumed to be “clean,” meaning not containing any outliers. Using this subset, a regression model for the variable of interest is fit. Deviations of the fitted values generated by the model from the actual sample values are then calculated, and a new “clean” subset is formed containing the observations that produce the $m + 1$ smallest distances. This procedure is repeated until the calculated deviations of the observations outside the clean subset are all considered too large, and therefore outliers, or until the subset is exhaustive.

Potter (1990) describes an outlier identification method that does not use actual survey data, but rather relies on an assumed distribution of the weights. In this procedure, a trimming level is prespecified based on a probability of occurrence according to the distribution model. For instance, if the trimming level is set to 1 percent, any observation whose probability according to the model is at most 0.01 will be considered an outlier and a candidate for trimming. Another option is to implement this procedure for an initial trimming and then distribute the excess weights among the untrimmed cases. The parameters of a new sampling weight distribution can then be estimated and a revised trimming level set. A second trimming can then take place followed by another redistribution of the excess weights to untrimmed cases.

Another popular outlier detection procedure involves examining the contribution of each sampling weight to the sum measure of entropy. Those weights that contribute substantially to the entropy are considered outliers. One such method, referred to by Potter as the NAEP procedure, identifies outliers based on the contribution to the overall sampling variance. This is accomplished by computing a value using the sum of the squared weights, $c \sum w_k^2 / n$, where c is a preset constant and n is the number of observations. For each observation, the squared weight is compared to the above quantity and those cases exceeding that trimming level are trimmed to the square root of that value. The excess weight is redistributed to untrimmed cases to retain the sum of the weights. The process is then repeated until no case remains with a squared weight that is larger than the trimming level. A similar method compares each sampling weight to some value, k , times the median of the sampling weights. The median is used instead of the mean because the mean can be heavily influenced by extreme weights. Often, k is set to be a simple constant such as 3 or 5, but it can also be defined by the distribution of $\{w_i / median, i = 1, 2, \dots, n\}$. The weights larger than the trimming level are trimmed to that value and the excess weights are redistributed among the untrimmed cases. To limit the number of cases to be trimmed, k can be increased.

The methods described above are the primary ones that we considered based on the literature review, although there are certainly several others available. Our review suggested that none of these methods as described in the literature could be automated and consistently identify observations that would be good candidates for trimming. Most of these procedures were not implemented in our study because some feature did not make them useful for this study, or because they did not apply well to household survey data. Several methods were designed to be more useful for identifying outliers in characteristics rather than outliers in survey weights. For example, the forward search method described by Chambers was designed to identify extreme observations (y -values), not extreme weights. An attempt was made to modify the procedure to handle the weights; however, even after several iterations, the largest deviations did not always correspond with the largest weights. Therefore, the “clean” subset could not be relied upon to include all nonoutlying cases. The weight distribution procedure specifically addresses outliers in weights, yet this procedure also was problematic. The method requires that the distribution of the weights be either known or accurately estimated. A few known distribution models including the ones suggested by Potter were tested to see if the household survey data in our work fit these models. However, no reasonable fit was found, and the method was not explored any further. With both the NAEP procedure and the k *median method, a constant is required to determine the trimming level. We examined some constants, but found that household survey weights vary

considerably across subgroups. No preset value seemed appropriate across household survey designs, and this limits that ability to automate the procedure and make it portable.

2.2 New Techniques

In addition to the methods found in the literature, other methods that were not specifically designed for sample surveys were reviewed and adopted for this purpose. The procedures that were deemed most appropriate were those that use the spacings between the weights as a means of identifying outliers (see David 1970). To implement these procedures, the weights are first ranked from largest to smallest. Using order statistic notation (i.e., $w_{(n)}$), where n is the number of observations), the four largest weights in the dataset from largest downward are: $w_{(n)}$, $w_{(n-1)}$, $w_{(n-2)}$, and $w_{(n-3)}$. A “spacing” is the distance between a ranked weight w_i and the next ranked weight $w_{(i-1)}$, i.e., the spacing $z_{(i)} = w_{(i)} - w_{(i-1)}$.

Two new methods for identifying largest weights in household surveys were developed using this concept of spacings. The first method aspires to identify large gaps in the weight distribution for the largest of the weights. For each weight, the spacing between it and the next largest weight is compared with the spacings between the next five pairs of ranked weights. The value

$$d5_space_{(i)} = \frac{z_{(i)}}{z_{(i-1)} + z_{(i-2)} + z_{(i-3)} + z_{(i-4)} + z_{(i-5)}}$$

increases when an observation is considerably larger than the next largest weight, in comparison to how much the next few weights vary from each other. This measure shows when there are large jumps in the distribution of the weights, which is an indicator that the weight is an outlier and should be considered for trimming. The second spacings method examines the distance between a weight and the next largest weight relative to the size of the weight. After some examination of an appropriate measure, we defined $rel_space_{(i)} = \frac{z_{(i)}}{w_{(i)}} \times 10$. This definition allows

for the procedure to be implemented in the same way for different groups, regardless of how the magnitude of the weights may differ across subgroups.

Another method closely related to the NAEP procedure described by Potter is proposed to measure the effect on the variance estimates by examining the effect of dropping a particular weight. This method, called the RV method, compares an estimate of the effective sample size, as a function of the relative variance, given that the i th weight is dropped. After several iterations, we decided on the formulation given below:

$$RV_{(i)} = \frac{\hat{Effss}_{(i)} - \hat{Effss}_{(i-1)}}{\hat{Effss}_{(i-1)} - \hat{Effss}_{(i-2)}},$$

where

$$\hat{Effss}_{(i)} = \frac{i}{1 + rel_var_{(i)}}.$$

For example, when calculating $RV_{(n)}$ of the largest weight $W_{(n)}$, $\widehat{Effss}_{(n)}$ will be calculated using all observations. The quantity $\widehat{Effss}_{(n-1)}$ will be calculated using $n - 1$ observations after dropping the n th or largest weight. To calculate $\widehat{Effss}_{(n-2)}$, $n - 2$ observations are used, after dropping the $n - 1$ and n th observations, or the two largest weights.

2.3 Composite Score

The new methods described above and the methods from the literature search were explored using a household survey dataset. No single measure was found that clearly identified the vast majority of cases that were deemed outliers without identifying far too many cases that should not have been classified as outliers. The strategy followed was to develop criteria for trimming based on a combination of the methods that were deemed useful for household survey data. During this process, a variation of the relative distance method was implemented. The measure is the spacing between the relative distance for a weight and the relative distance for the next largest weight. That is, $diff_dist_{(i)} = distance_{(i)} - distance_{(i-1)}$, where $distance_{(i)}$ was defined earlier as the relative distance for weight $w_{(i)}$.

Several different criteria were developed and tested, involving the three new measures ($diff_dist$, $d5_space$, and rel_space). There was not enough consistency in the behavior of RV to support including it as part of the criterion. Instead, it was treated as a source of additional information to help make decisions about cases that were questionable candidates for trimming.

The final procedure identified any observation meeting **all** the following criteria as a candidate for trimming:

- $diff_dist \geq 1.0$;
- $d5_space \geq 0.9$; and
- $rel_space \geq 1.0$.

For the observations meeting all three criteria listed above, a composite score was calculated. The composite score is the sum of the values of each of the three measures ($diff_dist$, $d5_space$, and rel_space). The score also includes a “penalty” that reduces the score as the number of cases to be trimmed increases. The rationale for the penalty is as follows. When any observation with a score above a certain level is considered for trimming, it also implies trimming all the weights that are ranked higher. To reduce the chances of trimming too many cases, a penalty of $\frac{n - rank}{2}$ (where rank is $n, n-1, n-2, \dots, 1$, for the weights from largest to smallest) is deducted from the initial score. In this way, there is little to no penalty for trimming a few cases and a larger penalty for trimming more.

After evaluating different scores based on data from two surveys discussed in the next section, a scale was devised to aid in making trimming decisions. Even though the goal was to develop a fully automated procedure, it became evident that in many situations, some additional scrutiny

may still be required to make a decision for the questionable cases. Thus, three levels associated with the scale score were proposed. They are:

- **Score > 8**–Automatic: these cases are considered definite outliers that should be trimmed.
- **Score between 4 and 8**–Questionable: these cases have extreme weights by at least some of the criteria, but before the decision to trim is made, further evaluation is needed. The RV measure may be useful along with visual review of graphs of weights.
- **Score < 4**–No Action: these cases generally should not be trimmed.

The scale and proposed cut-offs are examined in the next section.

3. Empirical Study

3.1 Study Design

To evaluate the proposed outlier detection methods for household surveys, we use data from two random digit dial (RDD) telephone surveys. The first survey is the National Survey of America's Families (NSAF) conducted by Westat for the Urban Institute to study status of families as changes in social programs were implemented beginning in the late 1990s. The survey collected information on the economic, health, and social dimensions of the well-being of children, adults under the age of 65, and their families in 13 states and the balance of the nation. There were three rounds of data collection: 1997, 1999, and 2002. For more information on NSAF, see the Urban Institute web site listed in the references that contains a variety of methodological reports on the survey design and weighting. The second survey is the California Health Interview Survey (CHIS), a collaborative project of the UCLA Center for Health Policy Research, the California Department of Health Services, and the Public Health Institute. In this survey Westat telephone interviewers collected information on if, where, and how people get health care in California. The sample was allocated by county and aggregates of smaller counties with supplemental samples of selected populations and cities to form 41 sampling strata. There have been two rounds of data collection: 2001 and 2003. For more information on CHIS, see the UCLA web site listed in the references that contains methodological reports for both surveys.

These two surveys were chosen for several reasons. First, the surveys exhibit differential selection probabilities that vary by strata or subgroups. Second, both surveys have the features of multiple weights (for different groups, such as adults and children). Third, both surveys have been conducted more than once. These features allow us more opportunity to evaluate the proposed methods under varying circumstances and, thus, better assess the portability of the methods.

As noted earlier, outliers due to extremely large weights in household surveys are typically due to large unequal probabilities of selection. These surveys also have multiple weighting adjustment factors, such as nonresponse and poststratification, but the sizes of these adjustments are usually controlled more by various choices in the weighting, such as the selection of nonresponse adjustment cells. The weights for the NSAF and CHIS are typical household surveys in this respect. To illustrate the variability in the probability of selection weights,

consider sampling children in CHIS. In this survey, if a household has children under age 12, one child is selected for the sample. Thus, if a household has seven children under age 12, the weighting factor for the selected child is 7, the inverse of the probability of selection. On the other hand, if a household has only one child under 12, then that child would have a factor of 1.

Differential sampling weights in CHIS are associated with the following selection probabilities:

- initial sampling by strata;
- oversampling by ethnic population; and
- sampling person within household.

Similar features are present in NSAF, where the probabilities of selection include differential factors associated with:

- the initial sampling by state;
- subsampling by income level;
- subsampling households without children; and
- sampling person within household.

3.2 Evaluation

In this section we describe the findings of the empirical study using data from NSAF and CHIS. As alluded to in earlier sections, the development of the methods of detecting outliers was exploratory in the sense that several approaches and measures were considered and evaluated before deciding on the specific ones. Using the same datasets to determine the procedures and evaluate them is a bit unusual. To ameliorate this problem, the exploratory work was done on a dataset from one survey, then evaluated using a dataset from the other survey. Thus, the methods are probably more robust than might have otherwise been true. Nevertheless, we consider all the developments exploratory and there is a definite need for more rigorous evaluation.

One of the key approaches to the study was to treat major subgroups from each survey separately. In the NSAF, 13 geographic areas have probabilities of selection that vary widely, and these areas are treated as separate subgroups in the analysis. In the CHIS, each of the 41 geographic areas (counties or groups of counties) is treated separately because the rates differ greatly across these strata. While this is a survey-specific aspect of the application, at least this level of adjusting to the specifics of the survey may be necessary to gain some portability across surveys.

Before presenting the findings, it may be useful to describe some of the motivation for the approaches we developed. The main motivation was our experience in reviewing graphical outputs to identify outliers from sample surveys. Invariably, graphs of the weights in a survey are revealing, and the gaps or spacings between the largest weights is a natural criterion considered in visually identifying outliers. If the gaps are large, then trimming may be needed.

One typical structure is that the largest weight is distinct in the graph and appears as an obvious outlier. This type of situation can be easily detected using most detection methods, including the

three criteria and scale score method we propose. However, even in this case it is not always easy to determine if the largest weight is so large that it should actually be trimmed. Consequently, in applied work the treatment of the weights may not be consistent. When there are several large weights in the tail of the weight distribution—a situation that entails more complexity than the single large weight case—the size of the spacings of the weights is very influential in determining which weights should be treated as outliers. In both cases, the development of automatic detection methods is useful.

Because of space limitations in this paper, we limit the graphs and tables presented here. One graph and two tables from the research were selected to illustrate several key concepts. The first graph is from the 1999 child survey in the balance of U.S. (the set of states that were not identified as a separate geographic sampling area) for the NSAF. A table from the same source is also presented. In addition, data from the 2003 CHIS adult survey for selected geographic strata are presented in a table.

Figure 1 shows the three measures for the 13 largest weights in the balance of U.S. sample from the NSAF. Note the measures jump up for the fifth largest weight ($n - 4$). This weight satisfies all three criteria listed in the previous section and, therefore, constitutes a potential cut point. This graph also demonstrates that it is possible for more than one weight to satisfy all three criteria; in this case the largest and fifth largest weight satisfy all three conditions. This implies that it is necessary to examine a relatively large number of individual weights (say 25) to find the last one in the series that meets all three criteria for a cut point. Using the three proposed criteria, the five largest weights are identified as potential outliers and considered candidates for trimming.

Once the candidate weights for trimming are identified, the scores for those that meet all three criteria are computed. Table 1 gives the scores for the seven largest weights in the sample for the balance of U.S. along with some other details not shown in figure 1. The *diff_dist* for the largest weight is 9.2, which is very large compared to the cut-off score, whereas the measure of *d5_space* is only 0.9. This example highlights why it is important to look at the three measures together to identify outliers consistently.

Figure 1 and the numbers in table 1 reveal why the three measures are so different for the largest weight in this sample. The value of *diff_dist* depends on the distribution of all the weights in the sample, while the measure of *d5_space* depends entirely on the six adjacent weights. As shown in table 1, there are a number of sizeable gaps among the largest weights, which is why the *d5_space* is relatively small in this case. The score for the largest weight and for the fifth largest weight ($12.4 > 8$) is in the automatic trim range. Thus, the five largest weights should be trimmed according to the proposed guidelines.

The table also contains other statistics that help explain the variation in weights. The variability of the weights is largely due to the sampling of one child from a household where some households have a large number of eligible children. The RV of the weights shown in the table are consistent with the three main criteria in this example. The RV for the largest and fifth largest weight are much greater than for the other weights. In other situations we found that this was not always the case, and the RV could provide independent information that could be used to help determine whether to trim weights that had scores in the range of 4 to 8. Despite our belief that

the RV is valuable, we have not been able to develop a reasonable criterion for using this measure directly in identifying outliers.

Table 2 shows summary data similar to that in table 1, but these data are from select strata from the 2003 CHIS adult file. The strata were chosen to include many different situations that were encountered in this study. Some strata had no weights that met all three criteria so no candidate outliers were detected. We have not included any of these strata in the tables. Below we review the issues in each stratum.

In Stratum 3, data are given for the six largest weights. Only the two largest weights meet the three criteria and the scores for both of them are greater than 8. Following the guidelines, these two points should be trimmed. In Stratum 4, the 10 largest weights are shown because both the largest weight and the ninth largest weight (rank = $n - 8$) satisfy all three criteria simultaneously. The nine largest weights are candidate outliers, but the score for the ninth weight is just negative, so only the largest weight should be trimmed because it is the only one with a score greater than 8. The next stratum, Stratum 7, has 11 observations listed in the table. In this situation the first weight that meets all three conditions is the 10th largest weight, and the score for that weight is small (< 4). Thus, no weight should be trimmed according to the guidelines. Note that in this stratum, the penalty had the effect of moving this point from the questionable category to the no action category. In Stratum 11, the largest weight is the only candidate outlier according to the three criteria, and it has a score of 6.6, which is in the questionable range. The RV criterion may be useful to inform the trimming decision. When we carefully examined this weight, we came to different conclusions about the need for trimming and believe that this may be contingent on the uses of the survey. The questionable range seemed appropriate for this weight. Stratum 15 shows a situation in which the largest weight should definitely be trimmed, and the second largest falls into the questionable range. In this stratum either one or two weights might be trimmed.

Figure 1. Three Measures in the Balance of U.S. Stratum from the 1999 Child NSAF File

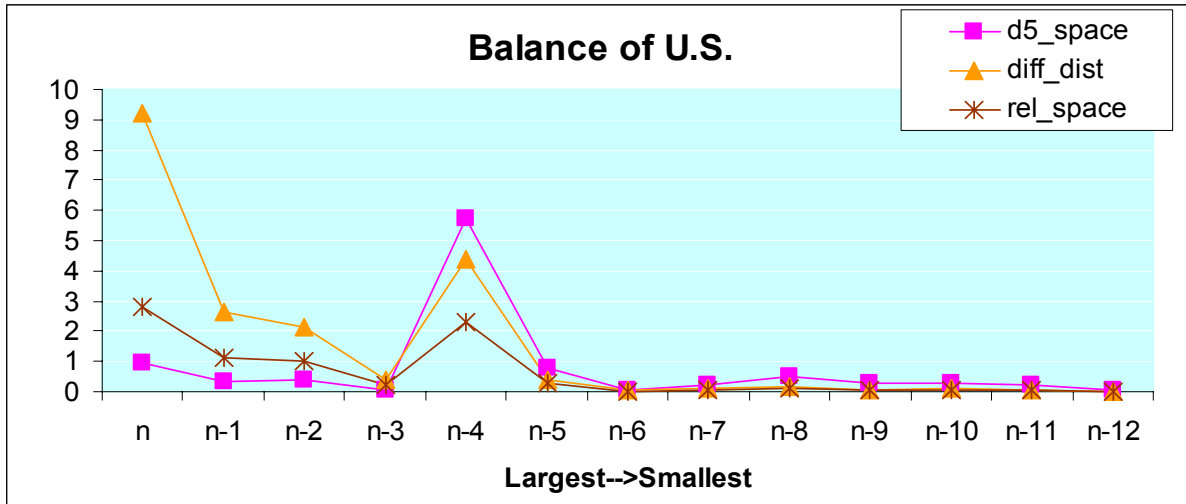


Table 1. Statistics from the Balance of U.S. Stratum from the 1999 Child NSAF file

Rank	Weight	# of kids	diff_dist	d5_space	rel_space	RV	Score
n	63,109	6	9.2	0.9	2.8	2.2	12.9
n - 1	45,528	7	2.7	0.4	1.1	1.4	--
n - 2	40,470	8	2.1	0.4	1.0	1.3	--
n - 3	36,423	6	0.4	0.1	0.2	1.1	--
n - 4	35,696	7	4.4	5.7	2.3	2.4	12.4
n - 5	27,381	2	0.4	0.8	0.3	1.1	--
n - 6	26,653	5	0.0	0.1	0.0	1.0	--

Note: Score is only computed for weights that satisfy all three criteria.

Table 2. Statistics from Selected Strata from the 2003 CHIS Adult File

Stratum	Rank	Weight	<i>diff_dist</i>	<i>d5_space</i>	<i>rel_space</i>	RV	Score
3	n	17,163	19.5	2.0	4.2	3.6	25.7
3	n - 1	9,877	7.2	2.7	2.7	2.3	12.1
3	n - 2	7,198	0.7	0.3	0.4	1.1	--
3	n - 3	6,940	0.2	0.1	0.1	1.0	--
3	n - 4	6,871	1.4	0.6	0.8	1.2	--
3	n - 5	6,339	0.4	0.2	0.2	1.1	--
4	n	6,970	4.9	1.2	2.0	1.8	8.1
4	n - 1	5,542	1.3	0.3	0.7	1.2	--
4	n - 2	5,167	0.7	0.2	0.4	1.1	--
4	n - 3	4,974	1.7	0.4	1.0	1.3	--
4	n - 4	4,484	0.0	0.0	0.0	1.0	--
4	n - 5	4,478	0.5	0.2	0.3	1.1	--
4	n - 6	4,321	1.7	0.7	1.2	1.4	--
4	n - 7	3,821	0.2	0.1	0.2	1.0	--
4	n - 8	3,763	1.4	1.1	1.1	1.4	-0.4
4	n - 9	3,364	0.0	0.0	0.0	1.0	--
7	n	4,740	4.8	0.4	0.8	1.2	--
7	n - 1	4,380	3.5	0.2	0.6	1.1	--
7	n - 2	4,118	0.1	0.0	0.0	1.0	--
7	n - 3	4,113	6.7	0.6	1.2	1.4	--
7	n - 4	3,613	0.1	0.0	0.0	1.0	--
7	n - 5	3,608	3.3	0.3	0.7	1.2	--
7	n - 6	3,360	4.5	0.5	1.0	1.3	--
7	n - 7	3,020	0.4	0.0	1.1	1.0	--
7	n - 8	2,992	2.0	0.3	0.5	1.1	--
7	n - 9	2,842	4.0	1.5	1.1	1.4	2.1
7	n - 10	2,539	0.6	0.2	0.2	1.0	--
11	n	3,799	3.0	1.7	1.9	1.9	6.6
11	n - 1	3,068	0.1	0.1	0.1	1.0	--
11	n - 2	3,033	0.7	0.4	0.5	1.2	--
15	n	6,749	2.3	0.7	5.4	1.9	8.4
15	n - 1	5,213	2.1	0.9	3.8	1.8	6.3
15	n - 2	4,136	0.2	0.1	0.2	1.0	--

Note: Score is only computed for weights that satisfy all three criteria.

4. Conclusion

Our goal was to develop a method for automatically detecting outliers due to large survey weights that should be trimmed for application to household surveys. In reviewing the literature, we noted that many outlier detection methods were constructed more for detecting unusual values of characteristics than for dealing with weights. The methods that did deal with the weights directly tended to be survey-specific and required a fair bit of customization to the particular survey situation. As a result, we examined other alternatives, based largely on spacings

of the weights. When we examined these methods, we observed that they had difficulty identifying outliers without including a number of observations that might not be outliers.

As a result, we proposed a composite method. First, three criteria based on different outlier detection procedures were established; candidate outliers are those cases with weights that meet all three criteria simultaneously. A summative score with a penalty that increases with the number of identified outliers is then computed for the candidate outliers. All the observations with scores greater than 8 are considered outliers that should be trimmed. Those with weights less than 4 should not be trimmed. Those weights with scores between 4 and 8 are in the questionable range and survey-specific goals or other measures such as RV may guide the trimming decision for these points.

This method does not satisfy our initial goal of having a fully automated system, but we believe it greatly reduces the burden for outlier detection and may be portable across many household surveys. Further work needs to be done to evaluate the proposed method in other surveys and perhaps refine it. An obvious need is revision of the RV procedure or the development of another measure that can address the effect of the outliers on the variances of the estimates. Another clear need is to better place these practical methods into a theoretical framework.

After the outliers are identified by methods such as those proposed here, the weights of the outliers are trimmed. One issue that we did not address is what to do with the other weights when the largest weights are trimmed. If the other weights are not adjusted, then the sum of the weights of the survey will be biased downward. In many household surveys, trimming is the last weighting adjustment before poststratification. If this is the case, then an option is to poststratify the trimmed weight to known population control totals. The excess weight that was trimmed is redistributed as part of the poststratification. If weight trimming is an intermediate step in the weighting process, then the method of dealing with the excess weight may be more difficult. The trimmed weights may or may not be redistributed to preserve the weighted totals, depending on the specifics of the survey.

References

- Chambers, R, A. Hentges, and X. Zhao. 2003. "Robust Automatic Methods for Outlier and Error Detection." *Journal of the Royal Statistical Society A* 167:323–39.
- David, H. 1970. *Order Statistics*. New York: John Wiley and Sons.
- Lee, H. 1995. "Outliers in Business Surveys." In *Business Survey Methods*. New York: John Wiley and Sons.
- Potter, F. 1988. "Survey of Procedures to Control Extreme Sampling Weights." In *Proceedings of the Survey Research Methods Section of the American Statistical Association* (453–58). Alexandria, VA: American Statistical Association.
- . "A Study of Procedures to Identify and Trim Extreme Sampling Weights." In *Proceedings of the Survey Research Methods Section of the American Statistical Association* (225–30). Alexandria, VA: American Statistical Association.
- The California Health Interview Survey. www.chis.ucla.edu
- The National Survey of America's Families. anf.urban.org/nsaf.