Evaluating and Improving Methods Used in the National Survey on Drug Use and Health

Joel Kennet and Joseph Gfroerer

EDITORS

DEPARTMENT OF HEALTH AND HUMAN SERVICES
Substance Abuse and Mental Health Services Administration
Office of Applied Studies

Acknowledgments

This report was prepared by the Office of Applied Studies (OAS), Substance Abuse and Mental Health Services Administration (SAMHSA), and by RTI International, a trade name of Research Triangle Institute, Research Triangle Park, North Carolina. Work by RTI was performed under Contract No. 283-98-9008. At SAMHSA, Joel Kennet and Joseph Gfroerer were editors of the volume. At RTI, Mary Ellen Marsden was task leader for production of the report, and Richard S. Straw edited the report with assistance from Jason Guder, Claudia M. Clark, K. Scott Chestnut, and Kathleen B. Mohar. Also at RTI, Diane G. Caudill prepared the graphics; Joyce Clay-Brooks and Debbie F. Bond formatted and word processed the report; and Pamela Couch Prevatt, Teresa F. Gurley, Kim Cone, David Belton, and Shari B. Lambert prepared its press and Web versions. Final report production was provided by Beatrice Rouse, Coleen Sanderson, and Jane Feldmann at SAMHSA.

Public Domain Notice

All material appearing in this report is in the public domain and may be reproduced or copied without permission from the Substance Abuse and Mental Health Services Administration. However, this publication may *not* be reproduced or distributed for a fee without specific, written authorization of the Office of Communications, SAMHSA, U.S. Department of Health and Human Services. Citation of the source is appreciated. Suggested citation:

> Kennet, J., & Gfroerer, J. (Eds.). (2005). *Evaluating and improving methods used in the National Survey on Drug Use and Health* (DHHS Publication No. SMA 05-4044, Methodology Series M-5). Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies.

Obtaining Additional Copies of Publication

Copies may be obtained, free of charge, from the National Clearinghouse for Alcohol and Drug Information (NCADI), a service of SAMHSA. Write or call NCADI at:

National Clearinghouse for Alcohol and Drug Information
P.O. Box 2345, Rockville, MD 20847-2345
1-301-468-2600, 1-800-729-6686, TDD 1-800-487-4889

Electronic Access to Publication

This publication can be accessed electronically through the Internet connections listed below:

http://www.samhsa.gov
http://www.oas.samhsa.gov

Originating Office

SAMHSA, Office of Applied Studies
1 Choke Cherry Road, Room 7-1047
Rockville, MD 20857

August 2005

Contents

Contents (continued)

Contents (continued)

Contents (continued)

Authors

Peggy Barker, Substance Abuse and Mental Health Services Administration
Paul P. Biemer, RTI International
Katherine R. Bowman, RTI International
Jim Brantley, RTI International
Rachel A. Caspar, RTI International
James R. Chromy, RTI International
Andrew Clarke, RTI International
Alexander J. Cowell, RTI International
David Cunningham, RTI International
Elizabeth Dean, RTI International
Antonieta Díaz, RTI International
Joe Eyerman, RTI International
Peter Frechtel, RTI International
Joseph Gfroerer, Substance Abuse and Mental Health Services Administration
Rebecca A. Granger, RTI International
Marjorie Hinsdale, RTI International
Michael Hubbard, University of North Carolina at Chapel Hill
Art Hughes, Substance Abuse and Mental Health Services Administration
B. Kathleen Jordan, RTI International
Joel Kennet, Substance Abuse and Mental Health Services Administration
Michael Langer, RTI International
Daniel Mamo, formerly with RTI International, now with United Guaranty Corporation
Peilan C. Martin, RTI International
Madeline E. McNeeley, RTI International
Martin Meyer, RTI International
Dawn Odom, formerly with RTI International, now with Inveresk
Dicy Painter, Substance Abuse and Mental Health Services Administration
Michael A. Penne, RTI International
Christina Salinas, RTI International
Jeanne Snodgrass, RTI International
Julie Stivers, RTI International
Douglas Wright, Substance Abuse and Mental Health Services Administration

1. Introduction

Joel Kennet and Joseph Gfroerer
Substance Abuse and Mental Health Services Administration

The National Survey on Drug Use and Health, or NSDUH as it is called throughout this volume, is the leading source of information on the prevalence and incidence of the use of alcohol, tobacco, and illicit drugs in the United States. It is currently administered to approximately 67,500 individuals annually, selected from the civilian, noninstitutionalized population of the United States, including Alaska and Hawaii. A survey of this size and importance is compelled to utilize the latest and best methodology over all facets of its operations. Given the sample size and the careful sampling procedures employed, NSDUH provides fertile soil for the testing and evaluation of new methodologies. Evaluation of NSDUH methodologies has been and continues to be an integral component of the project. This includes not only reviewing survey research literature and consulting with leading experts in the field, but also conducting specific methodological studies tailored to the particular issues and problems faced by this survey (Gfroerer, Eyerman, & Chromy, 2002; Turner, Lessler, & Gfroerer, 1992). This volume provides an assortment of chapters covering some of the recent methodological research and development in NSDUH, changes in data collection methods and instrument design, as well as advances in analytic techniques. As such, it is intended for readers interested in particular aspects of survey methodology and is a must-read for those with interests in analyzing data collected in recent years by NSDUH. Unless otherwise noted, data for all analyses and tables are drawn from NSDUH.

This introduction begins with a brief history and description of NSDUH. A more detailed account can be found in Gfroerer et al. (2002). Prior methodological research on NSDUH then is described, followed by an account of the major methodological developments that were implemented in 2002. Finally, each of the chapters and their authors are introduced.

NSDUH, formerly called the National Household Survey on Drug Abuse (NHSDA), among other names, was initiated in 1971 as a result of legislation enacted in 1970 that created the Commission on Marihuana and Drug Abuse. The commission was charged with reporting to the President and Congress on the drug use problem. As a result, the 1971 Nationwide Study of Beliefs, Information, and Experiences was designed in order to gather reliable data on the prevalence of marijuana use in the United States. This became the first in a series of data collections that, as the demand for data increased, along with the funding to meet those demands, grew in size and complexity from its initial sample size of about 3,000 to its current multistage area probability sample of nearly 70,000 persons throughout the 50 States and the District of Columbia. Since 1990, the survey has been fielded annually, and in 1999 it underwent a major redesign, as the administrative mode was upgraded from paper-and-pencil interviewing (PAPI) to its current computer-assisted interviewing (CAI) mode, which utilizes audio computer-assisted

self-interviewing (ACASI) for the more sensitive sections of the questionnaire. In 2002, further design changes were introduced, including the offer of an incentive and the adoption of the current name, in order to enhance the likelihood of participation and improve the accuracy of the data. The survey is sponsored by the Substance Abuse and Mental Health Services Administration (SAMHSA) and is currently conducted under contract by RTI International (a trade name of Research Triangle Institute).

NSDUH over the years has adopted methods that have been tested and shown to be effective in small methodological studies. For example, the 1999 conversion from PAPI to ACASI was carried out after a series of three field tests. The field tests demonstrated that the ACASI mode yielded higher levels of reporting of substance use, which presumably indicated increased accuracy (Office of Applied Studies [OAS], 2001). This is just one example. The full breadth of methodological research undertaken in relation to NSDUH is too large to describe here. Readers are referred to Turner et al. (1992) for a sampling of earlier work and to Gfroerer et al. (2002) for a thorough description of the research accompanying the 1999 redesign. This publication is intended to continue the NSDUH tradition of disseminating its more important methodological developments.

Although methodological improvements in NSDUH have been shown to increase data accuracy, reduce expenditures per completed interview, and occasionally increase the comfort of field interviewers (FIs) in the performance of their daily tasks, these gains do not come without costs. Practically any change that is made to a survey, be it in the sampling methods, protocols of participant contact, methods of data collection, instrumentation, or analytic methodology, has the potential to affect estimates and therefore disrupt trends. To return to the previous example, the shift from PAPI to CAI in 1999 undoubtedly increased the reporting of sensitive behaviors, which is believed to be a shift toward truthful reporting. The shift to CAI also reduced the proportion of missing data and the number of routing mistakes made by interviewers and respondents. However, it became practically impossible to compare data obtained in 1999 with those obtained in prior years of the survey.

Given the demand for trend data in a substantive area like drug abuse, one must exercise a great deal of caution when deciding upon methodological enhancements. Gains in data accuracy and other benefits must be weighed against the possible inability to evaluate trends and their relation to the numerous programs of substance abuse prevention and treatment that are in place in the transitional time period at local, State, and national levels. When there have been reasons to suspect that methodological changes would yield disruptions in trends, NSDUH has typically taken a cautious approach by running split-sample experiments, such as those in 1994 and 1999, in order to estimate the magnitude of the effects of method changes prior to a complete conversion to the new method. When change effects were expected to be small, or measurable without a split-sample study, the typical procedure has been to simply implement the change.

In some cases, changes in data collection methods have had unexpected effects on estimates, even after bona fide efforts have been made to anticipate them. This was the case in both 1999 and 2002. In anticipation of the effects of switching from PAPI to CAI administration, a large, split-sample study was designed and carried out in 1999 wherein some respondents received the new CAI instrument while others received the old PAPI instrument. The increase in sample size

that was necessitated by the overall redesign, coupled with the demands of the split-sample study, resulted in the hiring of many new interviewers. The unanticipated effect in this case was that newer interviewers were obtaining reports of substance use more frequently than experienced interviewers, possibly due to their stricter adherence to study protocols (Hughes, Chromy, Giacoletti, & Odom, 2001). The interviewer experience effect on prevalence was strong enough to make it impossible to estimate the effect of mode of administration, which in turn led to disruption of the ability to measure trends. A similar disruption occurred in 2002 when respondent incentives were introduced in NSDUH. Although the results of a 2001 experiment indicated that the incentive would have no appreciable impact on prevalence estimates, reality dictated otherwise. The SAMHSA reports presenting NSDUH's summary of findings in 2001 and 2002 (OAS, 2002b, 2002c, 2003) revealed increased prevalence estimates across the majority of substances queried in the survey. It was unfortunate that a split-sample study was not carried out in anticipation of these effects, although the costs of such a study would have been very large.

Chapter 2 of this volume, by Kennet, Gfroerer, Bowman, Martin, and Cunningham, describes the incentive and other methodological changes that were implemented in 2002, as well as their net effect on response rates and data collection costs. However, what is not covered in Chapter 2 is the unintended effect of these changes, which is the prevalence increases mentioned in the previous paragraph. The magnitude of these increases was sufficient to compel NSDUH staff to consider 2002 to be a new baseline for the measurement of trends. To date, work continues in the effort to tease out the effects of the individual factors responsible for these increases in reported drug use.

Chapter 3, by Painter, Wright, Chromy, Meyer, Granger, and Clarke, takes a closer look at the incentive introduced in the first quarter of 2002 and its effects on the activities of FIs. Using data recorded by interviewers on their visits to selected households, this chapter provides a summary description of the overall changes in respondent cooperative behavior. Readers wishing to know the specifics of how giving $30 cash payments to respondents actually reduced the cost of data collection in this complex, multistage survey will find this chapter useful and interesting.

A survey that collects data throughout the year across the entire Nation will inevitably at some point in time be affected by large-scale contextual events, whether they are of natural or human origin. NSDUH data were being collected in New York City and in Washington, DC, at the time of the terrorist strikes on the World Trade Center and the Pentagon. NSDUH staff reacted quickly and increased the sample size in the New York City area during the fourth quarter of 2001 in order to examine substance use prevalence in the aftermath of these events. The results of this study can be found in a report published by SAMHSA (OAS, 2002a). For survey researchers, this study also was of interest because of the effects these events had on response rates. Chapter 4, by McNeeley, Odom, Stivers, Frechtel, Langer, Brantley, Painter, and Gfroerer, summarizes the results of the response rate analysis.

Chapter 5 is the last chapter in this volume centered on data collection. Written by Chromy, Eyerman, Odom, McNeeley, and Hughes, this study takes a closer look at the FI experience effect mentioned earlier, which was first discovered in the 1999 data. As stated above, FI experience was found to correlate negatively with the likelihood that respondents would report substance use (Eyerman, Odom, Wu, & Butler, 2002; Hughes et al., 2001; Hughes, Chromy,

Giacoletti, & Odom, 2002). Prior work also had shown that the amount of experience that FIs had on the project was positively related to their likelihood of gaining respondent cooperation. This chapter links these two findings and estimates their relative influence on prevalence estimates using a two-stage conceptual model, which is tested by a series of logistic regressions.

The next three chapters deal with the instrumentation used in NSDUH and the methods employed to ensure data quality. Chapter 6, by Hinsdale, Díaz, Salinas, Snodgrass, and Kennet, explains in detail the procedures used to evaluate and revise the Spanish translation of the NSDUH instrument. Numerous examples are provided of the kinds of inadequacies that were uncovered in the review and the manner in which they were resolved. Perhaps more important, the chapter proposes a sequence of best practices for the processes of creating and updating Spanish translations of surveys.

Chapter 7, by Penne, Snodgrass, and Barker, investigates the use of audit trails–computer-generated descriptions of respondent behavior during the interview–which reveal such things as time spent on a given question or series of questions, whether the respondent backed up and changed an answer, or the location and time at which a respondent broke off an interview. Penne, et al. used these data to determine whether respondents were shortcutting during the ACASI portion of the interview, whether interview breakoffs were related to particular portions of the instrument or to particular FIs, and whether respondents' ability to go back and change answers affected substance use prevalence rates.

Chapter 8 also is an illustration of some of the many possibilities that are brought about by the use of computers in collecting survey data. Given the sensitive nature of the data that are collected in NSDUH, it is not surprising that a fair proportion of the respondents refuse to answer certain questions, or choose "don't know" as their response. One way to partially alleviate this problem is to administer probes, accompanied by assurances of the confidentiality of the data. These can be tailored to match the particular response chosen (refused vs. don't know) or, for that matter, any other prior information provided by a respondent. In this chapter, written by Caspar, Penne, and Dean, the authors examine the effects of the use of probes on reporting of substantive responses, consequent changes in prevalence rates, and the extent to which the probes were triggered by various segments of the NSDUH sample.

In Chapter 9, the volume begins to focus more on analytic methodology than on instrumentation, as Biemer, Jordan, Hubbard, and Wright describe the adaptation of item count methods to drug use prevalence estimation. Item count is a survey method in which respondents are not asked directly whether they have engaged in a sensitive behavior. Prevalence estimates are derived from responses to questions containing lists of behaviors in which the respondent indicates how many he or she has engaged in. The chapter begins with an account of the development of the item count questions that were used in the 2001 and 2002 surveys, then moves on to the statistical description of the analytic methods used, and finally, the results are presented. This chapter is noteworthy because of the painstaking attention to detail that was applied in the development and testing of this indirect estimation methodology.

The final chapter in this volume focuses on a very specific component of analysis: the measurement of income in NSDUH. Because income is an important correlate of substance use,

it is imperative that it be measured as accurately as possible. However, assessing the accuracy of the income measure is not as straightforward as one might expect. The most practical method for doing so involves comparing the income distribution obtained by NSDUH with those obtained by other surveys with similar samples. Chapter 10, by Cowell and Mamo, describes the process and results of comparing 1999 income as obtained by NSDUH with the distributions obtained by the Current Population Survey (CPS) and the Statistics of Income, a stratified national sample of all income tax returns filed.

In conclusion, this volume contains a collection of some recent methodological research carried out under the auspices of the NSDUH project. Publishing these studies periodically provides a resource for survey researchers wishing to catch up on the latest developments from this unique survey. Readers from a variety of backgrounds and perspectives will find these chapters interesting and informative and, it is hoped, useful in their own careers in survey methodological research, drug abuse prevention and treatment, and other areas.

References

Eyerman, J., Odom, D., Wu, S., & Butler, D. (2002). Nonresponse in the 1999 NHSDA. In J. Gfroerer, J. Eyerman, & J. Chromy (Eds.), *Redesigning an ongoing national household survey: Methodological issues* (pp. 23-51, DHHS Publication No. SMA 03-3768). Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies. [Available as a PDF at http://www.oas.samhsa.gov/nhsda/methods.cfm#Reports]

Gfroerer, J., Eyerman, J., & Chromy, J. (Eds.). (2002). *Redesigning an ongoing national household survey: Methodological issues* (DHHS Publication No. SMA 03-3768). Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies. [Available as a PDF at http://www.oas.samhsa.gov/nhsda/methods.cfm#Reports]

Hughes, A., Chromy, J., Giacoletti, K., & Odom, D. (2001, August). Impact of interviewer experience on drug use prevalence rates in the 1999 NHSDA. In *Proceedings of the American Statistical Association, Survey Research Methods Section* [CD-ROM]. Alexandria, VA: American Statistical Association.

Hughes, A., Chromy, J., Giacoletti, K., & Odom, D. (2002). Impact of interviewer experience on respondent reports of substance use. In J. Gfroerer, J. Eyerman, & J. Chromy (Eds.), *Redesigning an ongoing national household survey: Methodological issues* (pp. 161-184, DHHS Publication No. SMA 03-3768). Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies. [Available as a PDF at http://www.oas.samhsa.gov/nhsda/methods.cfm#Reports]

Office of Applied Studies. (2001). *Development of computer-assisted interviewing procedures for the National Household Survey on Drug Abuse* (DHHS Publication No. SMA 01-3514, Methodology Series M-3). Rockville, MD: Substance Abuse and Mental Health Services Administration. [Available at http://www.oas.samhsa.gov/nhsda/CompAssistInterview/toc.htm]

Office of Applied Studies. (2002a). *Impact of September 11, 2001 events on substance use and mental health* (DHHS Publication No. SMA 02-3729, Analytic Series A-18). Rockville, MD: Substance Abuse and Mental Health Services Administration. [Available at http://www.oas.samhsa.gov/analytic.htm]

Office of Applied Studies. (2002b). *Results from the 2001 National Household Survey on Drug Abuse: Volume I. Summary of national findings* (DHHS Publication No. SMA 02-3758, NHSDA Series H-17). Rockville, MD: Substance Abuse and Mental Health Services Administration. [Available at http://www.oas.samhsa.gov/p0000016.htm#Standard]

Office of Applied Studies. (2002c). *Results from the 2001 National Household Survey on Drug Abuse: Volume II. Technical appendices and selected data tables* (DHHS Publication No. SMA 02-3759, NHSDA Series H-18). Rockville, MD: Substance Abuse and Mental Health Services Administration. [Available at http://www.oas.samhsa.gov/p0000016.htm#Standard]

Office of Applied Studies. (2003). *Results from the 2002 National Survey on Drug Use and Health: National findings* (DHHS Publication No. SMA 03-3836, NSDUH Series H-22). Rockville, MD: Substance Abuse and Mental Health Services Administration. [Available at http://www.oas.samhsa.gov/p0000016.htm#Standard]

Turner, C. F., Lessler, J. T., & Gfroerer, J. C. (Eds.). (1992). *Survey measurement of drug use: Methodological studies* (DHHS Publication No. ADM 92-1929). Rockville, MD: National Institute on Drug Abuse.

2. Introduction of an Incentive and Its Effects on Response Rates and Costs in NSDUH

Joel Kennet and Joseph Gfroerer
Substance Abuse and Mental Health Services Administration

Katherine R. Bowman, Peilan C. Martin, and David Cunningham
RTI International

In 2002, the National Survey on Drug Use and Health (NSDUH) began offering respondents a $30 cash incentive for completing the questionnaire. This development occurred within the context of a name change[1] and other methodological improvements to the survey and was associated with significantly higher response rates. Moreover, the increased response rates were achieved in conjunction with a net decrease in costs incurred per completed interview. This chapter presents an analysis of response rate patterns by geographic and demographic characteristics, as well as interviewer characteristics, before and after the introduction of the incentive. Potential implications for other large-scale surveys also are discussed.

Survey Changes in 2002

Several methodological changes and improvements were made in the 2002 NSDUH that may have affected response rates.

Decision to Offer Incentives

In recent years, there has been a great deal of concern generated by the downward trend in response rates obtained in many surveys. Nonresponse to this survey was of particular concern because many of the drug-using behaviors measured have low prevalence in the population and could be severely biased if nonrespondents are more likely to be users than respondents are. Furthermore, wide variability in response rates was observed among States, and it became increasingly important to ascertain that such variability was not responsible for differences in estimated prevalence rates. Due to the importance of obtaining accurate estimates of prevalence and incidence of drug use and abuse, both at State and national levels, reducing nonresponse on the survey became a priority in spite of the fact that NSDUH response rates had not decreased as they had in other surveys in recent years. It was felt at the time that the best course of action was to anticipate a downward trend in response rates and take measures to prevent such a downturn in NSDUH. The idea of offering an incentive to respondents was consequently entertained, owing at least in part to the successes reported in other survey efforts (Berlin et al., 1992;

[1] Prior to 2002, the survey was called the National Household Survey on Drug Abuse (NHSDA). In this report, it is referred to as NSDUH, regardless of the survey year.

Goyder, 1987; Groves & Couper, 1998; Singer, 2001), in particular those obtained in health surveys (Bryant, Kovar, & Miller, 1975; Ezzati-Rice, White, Mosher, & Sanchez, 1995).

In 2001, an experiment was carried out to examine the effects of incentives on response rates and on survey costs in NSDUH. Incentives of $0, $20, and $40 were compared in a sample design that took prior local response rates and prevalence estimates into account. The results of the experiment were promising. Screening response rates were not noticeably affected by the incentive and averaged a little less than 92 percent (these are weighted percentages). Interview response rates, on the other hand, were 69.2 percent in the $0 control group, 78.8 percent in the $20 group, and 83.3 percent in the $40 incentive group. Moreover, the net cost per interview decreased with incentives, even when the incentive amount was included in the calculation. Cost per interview in the $20 group was 4.98 percent lower than the control, and in the $40 group, costs were 3.92 percent lower than the control. The cost savings were gained, for the most part, by interviewers spending less time and travel in trying to obtain cooperation from respondents. Analysis of the data obtained in the experiment indicated no effect of incentives on prevalence rates. As a result of these encouraging findings, it was decided that NSDUH would offer an incentive to respondents, beginning in the first quarter of 2002 data collection.

The decision to offer $30 was made for two reasons. First, in the incentive experiment, the majority of the gains in response rates were realized with the $20 incentive. Offering $40 did not appear to attract additional respondents in numbers at all in proportion to the percentage increase in the incentive. Second, where $40 was offered, there was much greater variability in the cost per interview than where $20 was offered. There was, therefore, uncertainty as to whether offering $40 across the entire sample would yield as much cost savings as was realized in the incentive experiment. Offering $30 across the board was expected to yield most of the gains in response rates that $40 had yielded in the experiment, and at the same time, the cost per interview reasonably could be expected to be reduced significantly.

Decision to Change the Name of the Survey

Although the decision to offer incentives was likely to be the most important development in the survey in some time, other changes were being implemented simultaneously. For several years, there had been discussion of the idea that response rates might improve if the survey name reflected a more general emphasis on health topics and did not refer to "abuse." At one point, focus groups were conducted in an attempt to address this issue. According to several field interviewers (FIs), the term "drug abuse" was a turnoff to some potential respondents, carrying strong negative connotations. Also, using "household" in the title may have been seen as redundant. In the early 1990s, the idea to change the name of the survey was brought up for formal consideration, but was turned down by the Administrator of the Substance Abuse and Mental Health Services Administration (SAMHSA) at the time because of the wide familiarity of policymakers and the public with the existing name. The idea did not resurface until 2000, after a major redesign of the survey from the paper-and-pencil interviewing (PAPI) mode to the computer-assisted interviewing (CAI) mode and expansion to State-based sampling had been achieved. The U.S. Department of Health and Human Services (DHHS) Secretary Donna Shalala decided at that point that the name should be changed. After considerable efforts to consult with policymakers, researchers, and survey field staff, SAMHSA decided that the new name would be

the National Survey on Drug Use and Health, or NSDUH, beginning in 2002. This happened to coincide with the commencement of the incentive.

Increased Adherence to Study Protocols

Analyses of the data from the 1999 CAI-PAPI split sample uncovered troubling interviewer effects on drug use prevalence estimates. In particular, interviewer experience was found to be associated with respondents' reports of substance use (Gfroerer, Eyerman, & Chromy, 2002). Consequently, in 2001, survey staff embarked on an effort to observe FIs at work in order to better understand their interaction with the instrumentation, the equipment (the survey is administered via computer-assisted personal interviewing [CAPI] and audio computer-assisted self-interviewing [ACASI], using a handheld device for screening and a laptop for interviews), and respondents. In order to accomplish this, a number of individuals with experience in the survey accompanied FIs as they carried out their tasks and took structured notes on FI behavior. Through this process, it was found that more than a few interviewers diverged from standard protocol to some extent. Most of these departures from protocol were not particularly serious, in that they probably did not bias responses in most cases. However, some behaviors were particularly egregious, such as failure to read the informed consent script or to provide the respondent a written description of the study. Moreover, it was found that experienced interviewers were more likely to break from protocol than those with less than a year of experience.

As a result of the observed breaches of protocol, training and retraining programs were redesigned over a period of time, undergoing continuous quality improvement. It should be noted that interviewer training for NSDUH is no minor undertaking. New-to-project interviewers receive a week of intensive, standardized training at one of several sites nationwide. Veteran interviewers, that is, all interviewers currently working on the project, every January undergo a 2 to 2½ day refresher training. The effects of the changes in training procedure quickly became evident. Further observations determined that in 2002 there was significant improvement in interviewers' adherence to study protocols. It remains unclear what effect, if any, improved protocol adherence had on response rates or on prevalence rates. The process is mentioned here in order to point out that the initiation of the incentive and the name change were not the only methodological developments that took place in NSDUH between 2001 and 2002. Efforts to improve protocol adherence were evolving and ongoing throughout the transition and continue to do so to date.

Summary of Survey Changes in 2002

The point of the previous discussion was to make clear that the introduction of a $30 incentive in the 2002 NSDUH did not occur while the remainder of the survey stood still. In fact, several developments were at least somewhat likely to have affected response rates simultaneously. As a result, the true effects of the incentive, as well as the effects of each of the other changes, cannot be reliably estimated from the data at hand. However, as noted earlier, other surveys have reported considerable success in their experimentation with incentives. Additionally, in the 2001 NSDUH incentive experiment, a $20 incentive yielded an interview response rate that was 9.6 percentage points higher than the rate obtained with no incentive, and the $40 incentive yielded a response rate that was 14.1 percentage points higher than the control

rate. Given the observed power of incentives and the lower relative salience of the other changes made, it seems reasonable to assert that changes in response rates between 2001 and 2002 *can* be attributed primarily to the incentive.

Method

To demonstrate the effects of the incentive and other changes, a comparison is presented of the response rates, by quarters, in the years before and after the incentive was introduced. These rates then are broken down by geographic area, population density, respondent age, gender, and race/ethnicity. Response rates also are examined with respect to the characteristics of interviewers, including their prior experience on the survey, race/ethnicity, and gender.

Before examining the results it is important to understand how each type of response rate was calculated. Screening response rates (SRRs) were the proportion of selected, eligible dwelling units that yielded a completed screener. Interview response rates (IRRs) were calculated as the number of completed interviews over the number of individuals within all households actually selected to participate. It was possible to have none, one, or two persons selected at each screened household. Finally, overall response rates (ORRs) were simply calculated by multiplying SRR by IRR. All of the reported rates are weighted to account for differing selection probabilities across age groups.

The cost-per-interview figures can be thought of as field costs of data collection, including time, travel, and miscellaneous expenses incurred by FIs. The most important thing to note is that the cost figures presented *include* the cost of the incentive in 2002.

Results

This section is primarily concerned with IRRs because SRRs were largely unaffected by the changes that were made to the survey procedures (see ***Table 2.1***). All changes reported throughout the section are actual differences between rates observed before and after the incentive was implemented, not relative changes, unless otherwise indicated.

Table 2.1 Distribution of Screening and Interview Dispositions, by Quarter, 2000-2002 NSDUH

	2000				2001				2002			
	Q1	**Q2**	**Q3**	**Q4**	**Q1**	**Q2**	**Q3**	**Q4**	**Q1**	**Q2**	**Q3**	**Q4**
Screening (%)												
Unable to Contact	2.60	2.93	2.17	2.32	2.37	2.19	2.73	2.73	2.72	2.71	2.59	2.33
Refusal	4.68	3.86	3.96	4.04	4.90	4.89	4.84	5.07	6.58	6.00	5.39	5.52
Other Incomplete	0.53	0.43	0.55	0.54	0.60	0.80	0.77	0.68	0.90	0.81	0.84	0.79
Complete	92.19	92.79	93.31	93.10	92.13	92.12	91.66	91.52	89.80	90.48	91.19	91.37
Interview (%)												
Unable to Contact	5.81	5.87	4.74	5.76	4.84	4.92	5.74	5.80	4.06	4.27	5.24	4.48
Refusal	16.28	16.24	15.38	15.60	15.67	16.11	16.70	17.55	12.25	13.56	13.08	14.23
Other Incomplete	4.23	4.72	4.82	4.83	4.89	4.93	4.86	4.69	2.91	3.53	3.85	4.26
Complete	73.68	73.17	75.06	73.82	74.60	74.04	72.70	71.97	80.78	78.64	77.83	77.03

Respondent and Environmental Characteristics

Table 2.1 provides the quarterly distribution of screening and interview dispositions from 2000 through 2002. The rows labeled "Complete" contain overall SRRs and IRRs on NSDUH. It is readily discernible that IRRs received a considerable boost from the incentive and other changes that were implemented in Quarter 1 2002. While SRRs remained relatively flat (93 percent in 2000; 92 percent in 2001; 91 percent in 2002), IRRs jumped between 2001 Quarter 4 and 2002 Quarter 1 from 71.97 to 80.78 percent. This was by far the largest quarterly change in the IRR over the 3-year period charted. Additionally, the IRR rate remained relatively high for the remaining three quarters of 2002, although it did drop slowly over that time period.

Table 2.2 contains quarterly and annual weighted IRRs by geographic region, population density, age, gender, and race/ethnicity from 2000 through 2002. Regional results are described first, followed in succession by each of the remaining environmental and respondent characteristics.

The incentive had slightly more effect on IRRs in the Northeast and West than in the Midwest or South, with gains of 9.88, 9.10, 7.83, and 8.73 percentage points, respectively, between Quarter 4 of 2001 and Quarter 1 of 2002. However, the Northeast and West did not hold onto their gains. By Quarter 4 of 2002, IRRs had dropped off in these two regions to a level nearly 5 percentage points lower than those in the Midwest and South regions, where the initial gains were better maintained.

The effects of the incentive and other changes varied slightly by population density. Comparing Quarter 4 of 2001 with Quarter 1 of 2002, metropolitan statistical areas (MSAs) with fewer than 1 million inhabitants had an IRR that was 10.24 percentage points higher after the incentive compared with an 8.32 percentage rate difference in larger metropolitan areas (1 million or more inhabitants), and an 8.04 percentage rate difference in non-MSAs. In other words, the incentive and other changes seemed to have their greatest effect in smaller metropolitan areas. However, non-MSAs seemed to maintain their rates through the remainder of 2002, dropping only about 2 percent, compared with about a 4 percent drop for both high and moderate density metropolitan areas. The general pattern, in which areas with the highest population density have the lowest response rates, and vice versa, did not seem to be affected by the incentive and other changes.

The effects of the incentive and other changes were most obvious in certain respondent age groups. Comparing differences between annual response rates for 2001 and 2002 (see ***Table 2.2***), 12 to 17 year olds were 7.81 percent more likely to complete an interview, 18 to 25 year olds were 9.65 percent, 26 to 34 year olds were 4.59 percent, 35 to 49 year olds were 6.57 percent, while those 50 or older were only 1.62 percent more likely to complete the interview after implementation of the incentive and other changes.

Females and males were equally enticed by the incentive and other changes. Again comparing differences in annual interview response rates for 2001 and 2002, both males and females were about 5 percent more likely to complete the 2002 NSDUH than the 2001 survey. As ***Table 2.2*** makes clear, the response rate for females was consistently higher than that for

Table 2.2 Quarterly and Annual Weighted Interview Response Rates, by Geographic Region, Population Density, Age, Gender, and Race/Ethnicity, 2000-2002 NSDUH

	2000					2001					2002				
Domain	**Q1**	**Q2**	**Q3**	**Q4**	**Annual**	**Q1**	**Q2**	**Q3**	**Q4**	**Annual**	**Q1**	**Q2**	**Q3**	**Q4**	**Annual**
Geographic Region															
Northeast	68.90	71.69	73.42	72.75	71.68	71.60	72.71	71.47	68.44	71.02	78.32	73.34	75.95	74.37	75.57
Midwest	73.51	71.85	75.09	72.47	73.23	73.73	73.90	72.56	72.84	73.25	80.67	80.81	79.34	79.29	80.01
South	77.28	75.30	77.38	75.62	76.38	75.91	74.86	73.81	73.27	74.44	82.00	80.32	78.95	78.75	79.99
West	72.41	72.48	72.79	73.05	72.68	75.90	74.04	72.19	71.98	73.51	81.08	78.19	76.15	74.07	77.33
Population Density															
MSA, ≥1,000,000	71.05	71.39	73.91	72.23	72.14	72.33	71.61	70.03	70.29	71.07	78.61	77.30	76.60	74.14	76.65
MSA, <1,000,000	74.32	72.65	75.26	73.97	74.05	75.31	77.01	72.97	71.67	74.13	81.91	79.07	78.55	77.89	79.31
Non-MSA	77.77	77.62	77.06	76.62	77.27	78.00	74.52	77.55	75.59	76.38	83.63	80.54	79.25	81.73	81.31
Age (Years)															
12-17	82.29	82.84	81.83	83.37	82.58	82.46	83.01	82.19	81.10	82.18	91.16	89.33	90.03	89.44	89.99
18-25	77.10	77.43	77.44	77.43	77.34	77.11	76.90	74.80	73.33	75.51	87.00	84.84	84.57	84.21	85.16
26-34	73.87	76.00	75.88	73.91	74.92	76.08	75.75	73.45	74.07	74.82	81.81	78.79	78.82	78.10	79.41
35-49	73.71	72.99	74.80	74.11	73.89	73.61	73.25	71.15	71.64	72.38	79.29	77.55	79.98	78.87	78.95
50+	69.28	67.66	71.79	69.43	69.53	71.48	70.19	69.93	68.18	69.92	75.78	73.29	68.40	68.98	71.54
Gender															
Male	72.60	71.70	73.18	73.21	72.68	74.43	72.03	70.42	70.89	71.92	79.90	77.49	76.10	74.76	77.06
Female	74.70	74.50	76.77	74.38	75.09	74.76	75.88	74.83	72.94	74.58	81.64	79.71	79.44	79.19	79.99
Race/Ethnicity															
Hispanic	78.30	76.89	78.67	77.97	77.95	79.48	79.34	79.65	76.68	78.78	83.27	79.94	80.79	79.93	80.93
Black	75.30	76.24	76.25	76.87	76.19	76.67	77.71	73.64	72.14	74.98	82.09	82.31	84.02	80.55	82.24
Other	72.88	72.15	74.35	72.83	73.04	73.68	72.75	71.49	71.30	72.29	80.26	77.84	76.37	76.04	77.64

MSA = metropolitan statistical area.

males from 2000 to 2002, and the incentive and other changes did not affect the relative propensity of females and males to participate.

For this analysis, race/ethnicity was divided into three categories: Hispanic, black, and other. Comparing Quarter 4 2001 with Quarter 1 2002, it appears that among these groups, blacks were most affected by the incentive, with an IRR increase of nearly 10 percentage points. The IRR among others (non-black, non-Hispanics) increased nearly 9 percentage points, while the rate among Hispanics increased by about 6½ percentage points. During the remainder of 2002, the rate among others appeared to decline a bit more than in the other two groups.

Final Disposition of Cases

Tables 2.3 and ***2.4*** provide breakdowns of the final disposition of eligible screenings and interviews, including reasons for refusal when they occurred. Interestingly, there appeared to have been more screening refusals in 2002 than in 2001 (***Table 2.3***), which accounted in large part for the lower completion rate. Among the reasons for screening refusals, "nothing in it for me" was the most popular and grew by nearly three quarters of a percentage point after the incentive was introduced. ***Table 2.4*** shows that the incentive was associated with a decline in nearly all types of interview noncompletions, especially refusals. Note that "nothing in it for me" dropped by about 1½ percentage points between 2001 and 2002.

Table 2.3 Weighted Final Disposition of Eligible Screenings, 2000-2002 NSDUH

Screening (%)	2000	2001	2002
Completed	92.84	91.86	90.72
No One at Home	1.82	1.90	2.02
Not Available	0.24	0.24	0.26
Refusal (sum of all refusal categories)	4.14	4.93	5.86
Nothing in it for me	*2.30*	*2.82*	*3.56*
No time	*0.67*	*0.79*	*0.81*
Government/surveys too invasive	*0.71*	*0.78*	*0.85*
Gatekeeper/household member won't allow participation	*0.02*	*0.03*	*0.05*
Confidentiality or survey legitimacy concerns	*0.13*	*0.13*	*0.22*
House too messy/too ill	*0.04*	*0.04*	*0.09*
Other	*0.26*	*0.33*	*0.27*
Missing	*0.01*	*0.00*	*0.00*
Denied Access	0.45	0.35	0.30
Mental/Physical Handicap	0.16	0.20	0.20
Spanish Language Barrier	0.05	0.09	0.05
Other Language Barrier	0.27	0.39	0.35
Electronic Screener Problem	0.00	0.00	0.00
Other Eligible	0.02	0.04	0.23

Table 2.4 Weighted Final Disposition of Eligible Interviews, 2000-2002 NSDUH

Interview (%)	2000	2001	2002
Completed	73.93	73.31	78.56
No One at Home	2.02	2.00	1.81
Not Available	3.52	3.30	2.71
Parent Refusal	0.88	0.92	0.55
Refusal (sum of all refusal categories)	14.99	15.60	12.73
Nothing in it for me	*6.47*	*7.06*	*5.52*
No time	*4.22*	*4.53*	*3.79*
Government/surveys too invasive	*1.95*	*1.88*	*1.46*
Gatekeeper/household member won't allow participation	*0.32*	*0.41*	*0.58*
Confidentiality or survey legitimacy concerns	*0.34*	*0.37*	*0.42*
House too messy/too ill	*0.32*	*0.28*	*0.28*
Other	*0.81*	*0.70*	*0.50*
Missing	*0.57*	*0.36*	*0.18*
Denied Access	0.01	0.03	0.00
Mental/Physical Handicap	2.57	2.43	1.75
Spanish Language Barrier	0.08	0.17	0.19
Other Language Barrier	1.06	1.30	1.09
Electronic Screener Problem	0.01	0.01	0.01
Other Eligible	0.92	0.94	0.61

Interviewer Characteristics

Next, response rates are examined with respect to interviewer characteristics. ***Table 2.5*** contains annual weighted response rates broken down by interviewer's prior experience on the survey, race/ethnicity, and gender. It appears that inexperienced interviewers benefited slightly more from the incentive. In 2001, interviewers with some experience achieved a 73.80 percent IRR, while those with no experience achieved 70.57 percent. In 2002, these figures were 78.78 and 76.39 percent for experienced and inexperienced interviewers, respectively. The gap between experienced and inexperienced interviewer performance in obtaining interviews at screened households closed by nearly 1 percentage point.

Interviewer race/ethnicity appeared to have some effect on IRRs. White interviewers experienced the largest increase in IRR after the incentive (5.83 percent), followed by black interviewers at 3.77 percent, others at 3.07 percent, and Hispanic interviewers improving by 1.20 percent. Black interviewers continued to achieve the highest IRRs before and after the introduction of the incentive. Male and female interviewers benefited equally from the incentive, both experiencing increases in IRRs from around 73 to 78.5 percent.

Costs per Interview

Finally, the costs of data collection were examined. The cost per interview dropped nearly 7 percent from 2001 to 2002, after adjusting for wage and mileage-rate differences between the 2 years (data not shown). Again, this difference includes the $30 incentive that was paid to all interview respondents in 2002. The amount of savings overall was somewhat larger than what

Table 2.5 Weighted Response Rates, by FI Characteristics, 2000-2002 NSDUH

Domain	2000			2001			2002		
	Screening	Interview	Overall	Screening	Interview	Overall	Screening	Interview	Overall
Prior NSDUH Experience									
Some	92.97	73.96	68.76	92.27	73.80	68.09	90.96	78.78	71.66
None	92.32	73.72	68.06	89.83	70.57	63.39	88.57	76.39	67.66
FI Race/ Ethnicity									
Hispanic	92.13	69.41	63.95	91.47	73.46	67.20	87.30	74.66	65.18
Black	92.47	79.00	73.05	91.53	75.90	69.47	90.59	79.67	72.17
White	93.07	73.91	68.79	92.00	73.02	67.18	91.05	78.85	71.80
Other	91.61	73.31	67.16	91.27	72.45	66.13	88.92	75.52	67.15
FI Gender									
Male	91.26	71.61	65.35	91.34	73.09	66.76	89.81	78.51	70.51
Female	93.35	74.62	69.66	92.05	73.37	67.54	91.02	78.57	71.51

FI = field interviewer.

might be expected from the results of the incentive experiment, although the experiment did not include a $30 condition.

Discussion

The $30 incentive, with possible help from the other changes that were introduced in January 2002, produced dramatic improvement in the number of eligible respondents who agreed to complete the NSDUH interview. Moreover, the increase in respondent cooperation was accompanied by a decrease in cost per completed interview. Clearly, the adoption of the incentive was beneficial to all involved, with the possible exception of the FIs, who required fewer hours to complete their assignments on the project and consequently received less pay.

From these analyses, it appears that incentives had their most pronounced effect among people between the ages of 12 and 25. Because these are known to be the years in which substance use and abuse are most prevalent and have their greatest incidence, it seems the incentive was well spent in terms of capturing the population of most interest. However, serious thought needs to be given to methods for attracting those older than 25. It could be the case that $30 was simply insufficient to attract people who have settled into careers and/or other more rewarding activities, such as child rearing or retirement, or it could be that these people did not participate for other reasons. These reasons will have to be investigated and addressed in order for NSDUH to optimally cover the aging baby boom generation and other cohorts.

Further study also is called for in explaining the regional differences in response rate patterns produced by the incentive. It is not at all clear at this time why the improved rates were not maintained in the Northeast and West regions as well as they were in the Midwest and the South. Observation of these trends over a longer time period may provide some insights. The same can be said about the differences observed with respect to population density. These questions also call for more complex statistical modeling.

Given the size and diversity of the NSDUH sample, it appears that the decision to offer incentives in other survey research programs that are having difficulty with response rates is an easy one. Response rates improved while costs declined. On the other hand, careful thought and experimentation should be directed toward determination of the amount to be offered and the timing of the implementation. Costs, respondent burden, target populations, and other factors need to be considered in these decisions. Moreover, it is strongly suggested that an evaluation component be planned along with any implementation of incentives. This would imply that the onset of the incentive offer should not coincide with other methodological changes, unless the changes are taken into account in a careful multifactorial design.

In hindsight, one could argue that the methodological changes that were implemented in the 2002 NSDUH were significant enough to merit a split-sample study to investigate possible effects on prevalence estimates. As pointed out in the 2002 summary of NSDUH findings (Office of Applied Studies, 2003), the incentive, name change, and other developments quite probably were responsible for the observed jumps in drug use prevalence and the consequent decision to consider that year's results as the new baseline for trend measurement. Unfortunately, the incentive field test that was run the prior year did not detect any significant changes in prevalence. In addition, the costs of running a split-sample experiment in 2002 that would take the name change and the incentive into account while using a sample large enough to detect effects on prevalence would have been prohibitive.

A final, additional consideration related to the offer of an incentive is falsification. Obviously, FIs who falsified interview cases in 2002 could pocket the $30 incentive originally intended for the respondent. However, here again, this analysis was unable to determine definitively whether falsification increased after the introduction of the incentive. NSDUH verification and data quality control procedures, much like the training improvements mentioned earlier, have been evolving continuously over the past few years. Therefore, it is impossible to tell whether any increase in the number of falsified cases resulted from improvements in detection capability or an actual increase in the amount of falsification among interviewers. Thus, the final recommendation for other surveys considering the implementation of incentives is to incorporate, a priori, a solid set of procedures for efficiently preventing, detecting, and correcting to the extent possible any activities of this sort.

References

Berlin, M., Mohadjer, L., Waksberg, J., Kolstad, A., Kirsch, I., Rock, D., & Yamamoto, K. (1992). An experiment in monetary incentives. In *Proceedings of the American Statistical Association, Survey Research Methods Section* (pp. 393-398). Alexandria, VA: American Statistical Association.

Bryant, E. E., Kovar, M. G., & Miller, H. (1975, October). *A study of the effect of remuneration upon response in the Health and Nutrition Examination Survey: United States* (Vital and Health Statistics: Series 2, Data Evaluation and Methods Research, No. 67, DHEW Publication No. HRA 76-1341). Rockville, MD: Health Resources Administration, National Center for Health Statistics [Available as a PDF at http://www.cdc.gov/nchs/about/major/nhanes/nh1rrm.htm and http://www.cdc.gov/nchs/data/series/sr_02/sr02_067.pdf]

Ezzati-Rice, T., White, A., Mosher, W., & Sanchez, M. (1995). Time, dollars and data: Succeeding with remuneration in health surveys. In *Seminar on new directions in statistical methodology* (pp. 225-255, OMB Statistical Working Paper No. 29, Vol. 2). Washington, DC: Office of Management and Budget.

Gfroerer, J., Eyerman, J., & Chromy, J. (Eds.). (2002). *Redesigning an ongoing national household survey: Methodological issues* (DHHS Publication No. SMA 03-3768). Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies. [Available as a PDF at http://www.oas.samhsa.gov/nhsda/methods.cfm#Reports]

Goyder, J. C. (1987). *The silent minority: Nonrespondents on sample surveys*. Boulder, CO: Westview Press.

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: Wiley.

Office of Applied Studies. (2003). *Results from the 2002 National Survey on Drug Use and Health: National findings* (DHHS Publication No. SMA 03-3836, NSDUH Series H-22). Rockville, MD: Substance Abuse and Mental Health Services Administration. [Available at http://www.oas.samhsa.gov/p0000016.htm#Standard]

Singer, E. (2001). The use of incentives to reduce nonresponse in household surveys. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 163-177). New York, NY: Wiley.

3. Analysis of NSDUH Record of Call Data to Study the Effects of a Respondent Incentive Payment

Dicy Painter and Douglas Wright
Substance Abuse and Mental Health Services Administration

James R. Chromy, Martin Meyer, Rebecca A. Granger, and Andrew Clarke
RTI International

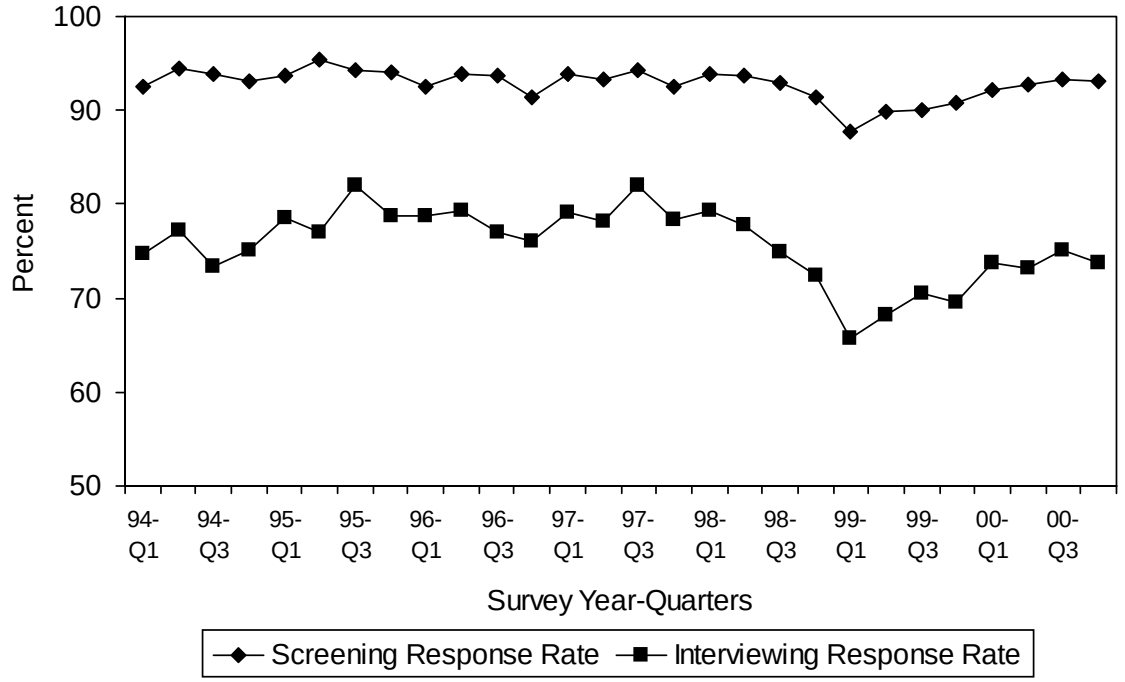
Background

The National Survey on Drug Use and Health (NSDUH) employs a multistage area probability sample to produce population estimates of the prevalence of substance use and other health-related issues. Letters are sent to selected households to alert potential respondents of the interviewer's future visit. Interviewers then visit the residence to conduct a brief screening, which determines whether none, one, or two persons are selected from the household. To maximize response rates, interviewers may make several visits to a household to obtain cooperation (the term "call" is used for "visits" from this point on with the understanding that this is a face-to-face survey; telephones are not used to contact potential respondents). In-person interviews are conducted with selected respondents using both computer-assisted personal interviewing (CAPI) and audio computer-assisted self-interviewing (ACASI). Sensitive questions, such as those on illicit drug use, are asked using the ACASI method to encourage honest reporting. A detailed description of the NSDUH methodology is provided elsewhere (RTI International, 2003).

During the late 1990s, NSDUH experienced a slight decline in response rates (see ***Figure 3.1***).[1] A closer examination of the data revealed stable noncontact patterns, but increasing refusal rates (Office of Applied Studies [OAS], 2001). This response rate decline was associated with the need to hire a large number of new interviewers for the expanded 1999 sample. Many of these new interviewers did not have the confidence or skills to overcome respondent refusals. Given the decline in response rates, NSDUH staff designed an experiment to evaluate the effectiveness of monetary incentives in improving respondent cooperation. A randomized, split-sample, experiment was conducted during the first 6 months of data collection in 2001. The experiment was designed to compare the impact of $20 and $40 incentive treatments with a $0 control group on measures of respondent cooperation and survey costs. The results showed that both the $20 and $40 incentives increased overall response rates while producing significant cost savings when compared with the $0 control group (Eyerman, Bowman, Butler, & Wright, 2002). Initial analysis showed no statistically detectable effects of the incentive on selected substance

[1] Prior to 2002, the survey was called the National Household Survey on Drug Abuse (NHSDA). In this report, it is referred to as NSDUH, regardless of the survey year.

Figure 3.1 Screening and Interviewing Response Rates, by Quarter, 1994-2000 NSDUH



use estimates. Subsequent analysis showed some positive and negative effects depending on the substance use measure when the $20 and $40 treatments were combined and compared with the $0 control group (Wright, Bowman, Butler, & Eyerman, 2002).

Based on the initial analysis of the 2001 experiment, NSDUH staff implemented a $30 incentive payment in 2002. This chapter analyzes the effect of the new $30 incentive on the data collection process as measured by record of calls (ROC) information. These records are created by the interviewers and describe the outcome of each visit to a particular household. The data contain extensive information on the amount of effort taken to contact and obtain cooperation from respondents. The effect of the incentives on response rates and costs is discussed by Kennet et al. in Chapter 2 of this volume.

Methodology

Study Populations

Two sets of study populations were available to examine the effects of incentives on the number of calls required to complete data collection processes. The first set consisted of the experimental treatment groups from the 2001 incentive experiment. This experiment was conducted in a sample of 251 of the 900 primary strata used in the 2001 survey.

The second set consisted of the annual samples for 2001 and 2002. The 251 primary strata included in the 2001 incentive experiment were excluded from the 2001 sample for these analyses; the weights used for the 2001 data were recalculated to reflect the additional sampling

and recalibrated to census demographic control totals. The entire sample was included in the 2002 incentive treatment group.

The 2001 incentive experiment data provided the best controlled comparisons in terms of reflecting the same time period and equating other field conditions, including the matching of interviewers on paired segments to which the incentive and control treatments were randomly applied. The second set of populations, 2001 non-experiment areas versus 2002, provided a much larger sample. Because initial analysis showed consistent results for the process data addressed in this chapter, only the results from the larger sample comparisons based on 2001 to 2002 comparisons are presented.

Record of Calls

Since the 1999 survey, a complete call history has been maintained for each selected dwelling unit using the handheld computer employed in the screening process.

Interviewers are instructed to record each attempted or actual contact with a selected dwelling unit. ***Figure 3.2***, taken from the field interviewer's (FI's) manual (RTI International, 2003), illustrates the computer screen display for recording each attempted screening call. The interviewer is instructed to record the result code and the method of contact; the computer automatically generates date and time information for each call record. ***Figure 3.2*** also illustrates a pop-up menu of possible result codes that may be selected by the interviewer for each screening call. When a screening is completed successfully and a roster of eligible persons has been obtained, the interviewer enters a command for the handheld computer to select the sample. The handheld computer automatically selects none, one, or two sample persons.

Figure 3.2 Screening Call Record as Seen by the Interviewers on Their Handheld Computers

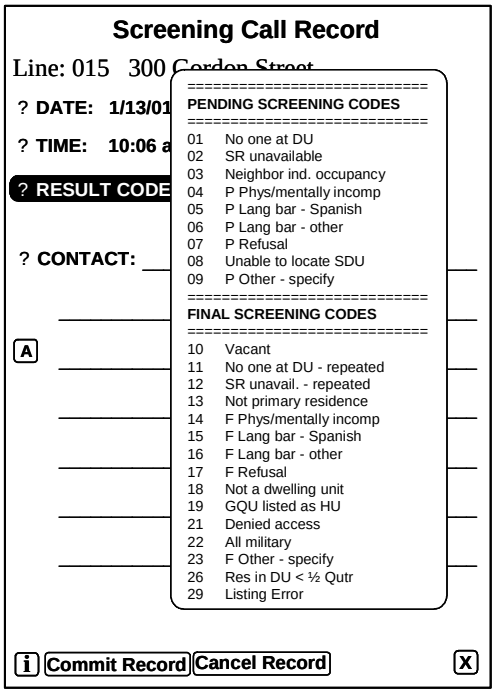


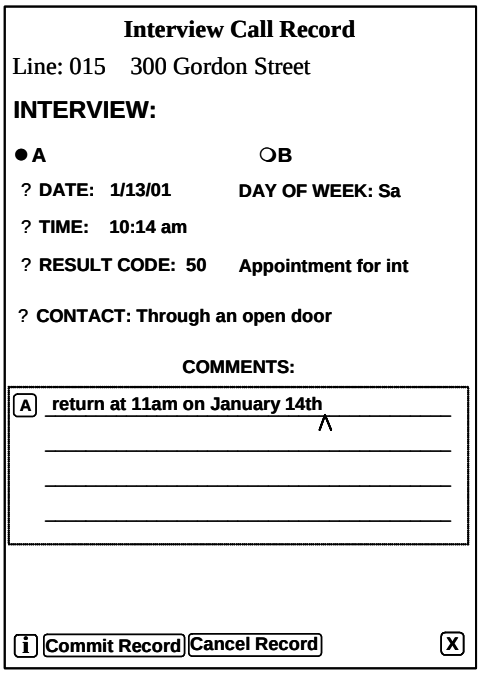
Figure 3.3 shows the computer display for entering the interview call records. If only one person is selected, that person is designated as person A. If two persons are selected, call records are maintained for both persons A and B. Note that it is common to record a screening event and one or two interviewing events on a single visit to a selected dwelling unit. On subsequent follow-up visits, events may be recorded for both respondents A and B.

On each day that an interviewer works, he or she transmits data to a central computer where the ROC becomes part of a larger field control system database. This control system is also used by supervisory staff to assign or reassign cases, to finalize pending cases, and to monitor and

control the progress of fieldwork. As a result, the control system contains many event records other than those relating to calls conducted by the FI. ***Table 3.1*** shows the screening event codes that were treated as a visit to attempt or complete screening at a selected dwelling unit in 2001 and 2002. ***Table 3.1*** categorizes the screening codes into pending and final categories. Final screening codes are grouped further into those indicating an ineligible sample dwelling unit, those indicating an eligible sample dwelling unit where screening was not completed successfully, or those for which screening was completed successfully at an eligible sample dwelling unit. For a successfully completed screening interview, the handheld computer automatically entered the final screening event code for no one selected (code 30), one person selected (code 31), or two persons selected (code 32).

Table 3.2 shows the event codes assigned during interviewing. Because the interview is conducted using a separate laptop computer, all interviewing codes (including the code for completed interviews) must be entered by the interviewer at or after the time of the visit. Interviewing event codes in ***Table 3.2*** are classified as pending and final codes. Final codes are classified as interview not obtained versus interview obtained. Final decisions about the usability of partially completed and completed interviews are made during the editing process based on minimum item response criteria (see Kroutil & Myers, 2002).

Figure 3.3 Interview Call Record as Seen by the Interviewers on Their Handheld Computers



For the purposes of studying the screening and interviewing process in this chapter, the initial categorizations in ***Tables 3.1 and 3.2*** are applied.

Calls and Call Days

Interviewers are instructed to work within an area segment conducting screening and interviewing for at least 4 hours each day they visit the segment. Unless they are able to complete some screening or interviewing successfully, they may visit the same dwelling unit several times during the same day and record an event code for each visit. Because repeated calls on the same day to try to establish contact with a selected dwelling unit or a previously selected person do not necessarily represent a significant and effective additional effort, the concept of a *call day* was used as an alternative measure. A call day for a sample dwelling unit was defined as any day on which one or more calls were made to a selected dwelling unit for any combination of screening and/or interviewing activity.

Analytic Approach

The general approach taken in this chapter to assess the impact of incentives on the screening and interviewing processes was to compare the distribution of the sample by the number of call

Table 3.1 Event Codes Assigned During Screening

Pending Screening Codes	
01	No One at Dwelling Unit
02	Screening Respondent Unavailable
03	Neighbor Indicates Occupancy
04	Physically/Mentally Incompetent
05	Language Barrier (Spanish)
06	Language Barrier (Other)
07	Refusal to Screening Questions
08	Unable to Locate Sample Dwelling Unit
09	Other
Final Screening Codes: Ineligible Sample Dwelling Units	
10	Vacant
13	Not a Primary Residence
18	Not a Dwelling Unit
19	Group Quarters Unit Listed as Housing Unit
20	Housing Unit Listed as Group Quarters Unit
22	Dwelling Unit Contains Only Military Personnel
25	No Eligible Sample Dwelling Unit Members
26	In Dwelling Unit Less Than ½ of the Quarter
29	Listing Error
Final Screening Codes: Screening Not Obtained	
11	No One at Dwelling Unit After Repeated Visits
12	Screening Respondent Unavailable After Repeated Visits
14	Physically/Mentally Incompetent
15	Language Barrier (Spanish)
16	Language Barrier (Other)
17	Refusal
21	Denied Access to Building/Complex
23	Other
Final Screening Codes: Screening Completed	
30	No One Selected for Interview
31	One Selected for Interview
32	Two Selected for Interview

Table 3.2 Event Codes Assigned During Interviewing

Pending Interview Codes	
50	Appointment for Interview
51	No One at Dwelling Unit
52	Respondent Unavailable
53	Breakoff (Partial Interview)
54	Physically/Mentally Incompetent
55	Language Barrier (Spanish)
56	Language Barrier (Other)
57	Refusal (by Respondent)
58	Parental Refusal for 12 to 17 Year Old
59	Other
Final Interview Codes: Interview Not Obtained	
71	No One at Home After Repeated Visits
72	Respondent Unavailable
73	Breakoff (Partial Interview)
74	Physically/Mentally Incompetent
75	Language Barrier (Spanish)
76	Language Barrier (Other)
77	Final Refusal by Respondent
78	Parental Refusal for 12 to 17 Year Old
79	Other
Final Screening Code: Interview Obtained	
70	Interview Complete

days required to finalize the sample dwelling unit status during the eligibility determination, screening, and interviewing processes. It is first necessary to determine whether a sample dwelling unit is eligible for screening. Only eligible sample dwelling units then are subject to finalization of the screening process. At the next step, only successfully screened dwelling units at which one or more persons are selected are followed into the interviewing process. If a sample dwelling unit is not eligible, the data collection process terminates very early. If screening cannot be completed or if no persons are selected, then the data collection process also terminates. The initial exploratory analysis showed that the number of calls required also depended on the final outcome; as an example, a successful interview will usually require fewer repeat visits than a case that was eventually finalized in the "no contact" category (codes 11 and 12 in ***Table 3.1***). The call day concept is used for most of the analysis, but data on calls also are used to simplify

the analysis of the interviewing process when one or two persons may be selected at a sample dwelling unit.

Each of the following data tables (***Tables 3.3*** to ***3.6***) shows the sample distribution and weighted[2] sample percentages. Upper categories are collapsed. The mean number of calls or call days is not presented because it is likely to be heavily influenced by a small number of extreme counts. These extremes are probably more closely related to the convenient travel patterns of the interviewer than to the incentive process applied. The bottom section of each data table shows the results of the tests of independence between the number of call days (or calls) required and the incentive treatment (2001 non-incentive areas vs. 2002 all incentive areas).[3] Because of the large sample sizes, statistically significant results need to be judged further to determine whether the differences noted are meaningful. Note that sample sizes for the 2001 non-incentive population are smaller than the sample sizes for the 2002 population because the primary strata in which the 2001 incentive experiment was conducted were excluded from these analyses.

Table 3.3 Screening Call Days, by Eligibility Category: 2002 (All Incentive) Versus 2001 (Non-Incentive) NSDUH

Number of Call Days	Sample Size		Weighted Percent	
	2002	2001	2002	2001
	Not a Residence			
1 day	10,948	8,501	49.2	45.8
2 days	5,240	4,423	22.6	23.0
3 days	2,479	2,001	10.9	11.0
4 days	1,260	1,115	5.7	6.3
5 to 9	2,129	1,796	9.8	11.0
10 plus	420	437	1.8	2.9
Total	22,476	18,273	100.0	100.0
	Residence, But Not Eligible			
1 day	1,964	1,799	36.8	35.2
2 days	1,200	1,062	22.4	22.0
3 days	650	574	11.9	11.6
4 days	454	386	8.9	7.7
5 to 9	854	903	15.7	18.2
10 plus	252	279	4.3	5.4
Total	5,374	5,003	100.0	100.0
	Eligible Dwelling Units			
1 day	59,201	48,365	39.2	39.2
2 days	32,195	25,756	21.3	21.0
3 days	18,654	14,631	12.6	12.0
4 days	11,820	9,526	8.0	7.8
5 to 9	22,071	17,962	14.9	15.3
10 plus	6,221	5,557	4.1	4.8
Total	150,162	121,797	100.0	100.0
All	178,012	145,073		
LLCHISQ tests of no interaction				*p*
Not a residence				0.0107
Residence, but not eligible				0.1696
Eligible dwelling units				0.0005

Results

Effect of Incentives on the Screening Process

Table 3.3 summarizes the number of screening call days required by dwelling unit eligibility. Nonresidential dwelling units can generally be identified quickly. Residences that are not eligible require more personal contact. The call days shown for eligible dwelling units represent all screening call days required to finalize the screening process. Although the distributions by incentive treatment (year) for the "not a residence" category are shown to be different by statistical testing, the differences in the distributions are quite small. This is consistent with

[2] The weights applied to both 2001 and 2002 data allow inference to the entire NSDUH target population. The standard analytic weights were used for 2002. Because data from the 251 primary strata used in the 2001 incentive experiment were excluded from the comparisons, a special weight based on an adjustment of the 2001 analytic weights was developed. The adjustment to the 2001 analytic weights accounted for removal of the incentive experiment subsample and for recalibration of the remaining sample to known totals.

[3] The statistical test presented is LLCHISQ option in SUDAAN (RTI, 2001, p. 279), which tests for no interaction in the log-linear model fit to the estimated cell proportions.

the expectation that incentives should have little effect at this stage of the data collection process. The differences in the call distributions for eligible dwelling units include further screening and interviewing processes, which are explored more fully in subsequent tables.

Table 3.4 provides more details on the number of screening calls required by final screening outcome for the "eligible dwelling units" category included in ***Table 3.3***. Note that over 40 percent of all completed screenings are completed on the first call day with or without incentives. Statistically significant differences in screening call day distributions are identified for the final categories of "language barrier," "refusals," and "screening respondents." For "refusals" and "screening respondents," the differences favor the 2002 incentive sample in that more screening interviews are finalized on the first or second call day and fewer screening interviews require 10 or more call days. The results for finalizing "language barrier" cases appear to favor the 2001 non-incentive sample, but are based on a much smaller sample.

In summary, the introduction of incentives had only small effects on the screening process, and most of them favored the use of incentives.

Table 3.4 Screening Call Days for Eligible Dwelling Units, by Final Screening Status: 2002 (All Incentive) Versus 2001 (Non-Incentive) NSDUH

Number of Call Days	Sample Sizes		Weighted Percent	
	2002	2001	2002	2001
	No Contact			
1 day	188	89	3.1	5.3
2 days	297	165	5.7	5.7
3 days	277	108	5.3	3.3
4 days	268	169	5.5	5.1
5 to 9	1,727	1,041	37.4	34.3
10 plus	1,973	1,427	43.1	46.5
Total	4,730	2,999	100.0	100.0
	Language Barrier			
1 day	19	20	3.7	4.9
2 days	130	172	24.8	39.1
3 days	107	76	20.1	15.6
4 days	79	55	15.6	11.5
5 to 9	158	118	29.8	23.4
10 plus	34	20	6.1	5.5
Total	527	461	100.0	100.0
	Refusals			
1 day	63	17	0.7	0.3
2 days	523	297	5.6	5.0
3 days	1,008	565	11.5	9.3
4 days	1,231	734	14.3	12.8
5 to 9	4,189	2,725	50.6	50.4
10 plus	1,542	1,208	17.4	22.1
Total	8,556	5,546	100.0	100.0
	Screening Respondents			
1 day	58,931	48,239	43.0	42.3
2 days	31,245	25,122	22.8	22.1
3 days	17,262	13,882	12.8	12.4
4 days	10,242	8,568	7.6	7.6
5 to 9	15,997	14,078	11.8	12.9
10 plus	2,672	2,902	1.9	2.7
Total	136,349	112,791	100.0	100.0

LLCHISQ tests of no interaction	*p*
No contact	0.1054
Language barrier	0.0389
Refusals	0.0002
Screening respondents	0.0000

Effect of Incentives on the Interviewing Process

The effect of incentives on the interviewing process was studied both in terms of the number of call days and the distribution of calls. Recall that when a screening interview resulted in the selection of sample persons, a screening event code was automatically entered on the handheld computer and the interviewer also entered an interview event code for each selected person. When looking at call days, this meant that no additional call days were required in order to record one (or two) interview event codes. Furthermore, if the interviews could be completed the same day, then no extra call days were required. This is a subset of the dwelling units where screening was successfully completed, as shown in ***Table 3.4***. ***Table 3.5*** shows the distribution of

additional interviewing call days by incentive treatment for dwelling units where one or two persons were selected.[4]

The distribution of the sample by additional interviewing call days clearly favors the use of incentives in obtaining early resolution of interviewing cases. When only one person was selected, no additional call days were required to complete interviewing 43 percent of the time with incentives, but only 36 percent of the time with no incentives. In addition, only about 2 percent required 10 or more additional call days with the use of incentives, while over 5 percent required 10 or more additional call days without incentives.

As expected, more additional call days were required when two persons were selected, but the results clearly favored the incentive treatment. Almost 25 percent of cases were resolved on the date of the screening interview even when two persons were selected; just under 17 percent of cases achieved this goal with no incentives.

Table 3.5 does not distinguish among cases by interview outcome. To examine the interviewing process by interviewing outcome, the concept of a call, rather than a call day, was implemented in ***Table 3.6***. At the completion of a successful screening with persons selected, the interviewer updated the ROC for each selected person. One recorded event for each respondent coincided with the completion of screening. Events may have been entered for both persons on subsequent visits also until at least one person's interviewing status was finalized. ***Table 3.6*** also is different from ***Table 3.5*** in that the sample units counted are sample persons in ***Table 3.6*** and sample dwelling units in ***Table 3.5***. No attempt is made in ***Table 3.6*** to identify which sample persons are from dwelling units with one or two persons selected.

Table 3.5 Additional Interviewing Call Days for Dwelling Units, by Persons Selected: 2002 (All Incentive) Versus 2001 (Non-Incentive) NSDUH

Number of Extra Call Days	Sample Sizes		Weighted Percent	
	2002	2001	2002	2001
	One Person Selected			
None	12,893	11,248	43.0	35.5
1 day	8,947	7,215	28.2	22.5
2 days	2,958	3,169	9.3	10.2
3 days	1,807	2,235	5.9	7.1
4 days	1,223	1,724	4.0	5.3
5 to 9	2,374	4,147	7.9	13.9
10 plus	536	1,565	1.7	5.5
Total	30,738	31,303	100.0	100.0
	Two Persons Selected			
None	6,181	2,873	24.8	16.7
1 day	8,041	3,626	31.3	20.9
2 days	3,639	2,134	14.4	12.5
3 days	2,091	1,608	8.4	9.6
4 days	1,427	1,250	6.0	7.7
5 to 9	2,990	3,586	12.4	22.3
10 plus	685	1,634	2.7	10.3
Total	25,054	16,711	100.0	100.0
LLCHISQ tests of no interaction				*p*
One person selected				0.0000
Two persons selected				0.0000

Statistically significant interactions of incentives with the number of calls required are noted for all outcomes except the "not competent to complete" outcome. The "no contact" cases were

[4] ***Table 3.5*** shows that a larger proportion of sample dwelling units had two persons selected in 2002 compared with 2001. This was caused by a change in selection algorithm implemented in 2002. The algorithm change preserved the person probabilities while increasing the numbers of person pairs in the sample. For more details, see Chromy and Penne (2002).

Table 3.6 Calls to Complete Interviewing, by Final Interview Status: 2002 (All Incentive) Versus 2001 (Non-Incentive) NSDUH

Number of Calls	Sample Sizes		Weighted Percent	
	2002	2001	2002	2001
	No Contact			
1 call	42	48	1.0	0.8
2 calls	95	70	2.8	1.4
3 calls	184	91	3.8	1.9
4 calls	625	357	17.2	9.3
5 to 9	1,361	1,239	40.4	31.1
10 plus	1,192	2,013	34.9	55.4
Total	3,499	3,818	100.0	100.0
	Not Competent to Complete			
1 call	756	688	61.9	60.9
2 calls	177	164	14.9	15.0
3 calls	83	101	7.6	8.2
4 calls	50	61	4.2	4.2
5 to 9	107	120	9.6	8.6
10 plus	28	62	1.8	3.3
Total	1,201	1,196	100.0	100.0
	Refusal, Breakoff, Other Unexplained			
1 call	764	571	10.5	6.7
2 calls	1,101	853	14.0	8.9
3 calls	1,061	1,000	14.2	10.3
4 calls	932	1,027	12.5	9.3
5 to 9	2,606	3,447	34.4	37.1
10 plus	1,146	2,740	14.5	27.9
Total	7,610	9,638	100.0	100.0
	Respondent			
1 call	24,345	14,540	35.7	30.2
2 calls	24,342	15,251	34.7	30.5
3 calls	8,003	5,871	11.6	11.6
4 calls	3,844	3,405	5.7	6.6
5 to 9	6,036	7,228	9.5	14.3
10 plus	1,701	3,405	2.8	6.9
Total	68,271	49,700	100.0	100.0

LLCHISQ	*p*
Interview calls by incentive treatment	
No contact	0.0000
Not competent to complete	0.6749
Refusal, breakoff, other unexplained	0.0000
Respondent	0.0000

finalized sooner with incentives in 2002, possibly because other pending cases were finalized sooner and additional trips to attempt contacts with only one or two persons did not appear justified; note that the number of "no contact" cases comes to about one person for every two area segments in both years. Somewhat surprisingly, the persons in the "refusal, breakoff, and other unexplained" category also were finalized more quickly with incentives. Persons who eventually responded did so on the first interviewing call about 36 percent of the time with incentives and about 30 percent of the time without incentives; recall that the first interviewing call in ***Table 3.6*** equates with no extra calls to complete this person's interview following the ***Table 3.5*** "extra call day" concept. The second call also was more productive with incentives perhaps indicating more willing participation by a selected person who was not present at the

time of the screening interview when incentives were offered. Respondents requiring 10 or more calls constituted less than 3 percent of final respondents with incentives and almost 7 percent of final respondents without incentives. This could be the result of individual cases being finalized with fewer calls overall, or the interviewers meeting their response goals earlier, making additional attempts unjustified.

Effect of Incentives on Response Demographics

Only limited analysis was conducted to assess the effect of the number of calls on sample demographics. ***Table 3.7*** shows the distribution of the sample by gender depending on the order number of the successful interviewing call. In either case, males responded in lower proportions to the early calls with the percentage male generally increasing slowly (or decreasing only slightly) with the call order. With incentives, the percentage male was slightly higher on the first call and increased more steadily with increasing calls.

Table 3.7 Percentage Male, by Call Order: 2002 (All Incentive) Versus 2001 (Non-Incentive) NSDUH

Call Order	2002	2001
1st call	46.01	45.96
2nd call	48.47	47.36
3rd call	49.12	49.53
4th call	50.05	49.40
5th to 9th calls	51.56	50.79
10th or later calls	51.97	50.07

Summary and Conclusions

NSDUH data show that the increases in response rates that accompany the use of incentives (Eyerman et al., 2002; Kennet, Gfroerer, Bowman, Martin, & Cunningham, 2003) also are accompanied by the need for fewer call days and fewer calls to finalize sample dwelling units and sample persons. Only small effects were noted for the call days associated with the screening process, with substantial effects noted for the interviewing process. The need for fewer visits to complete screening and interviewing also helps to explain reduced costs (OAS, 2002).

A quick look at demographic data by call order shows that the sample distribution on demographic measures could be changed by prematurely cutting off the follow-up of pending cases. More study is needed to determine what effect such policies might have on the principal study measures on substance use addressed in NSDUH.

The monitoring of ROC data provides a useful tool for ensuring that adequate follow-up procedures are being used within the limits of reasonable cost management.

References

Chromy, J. R., & Penne, M. (2002). Pair sampling in household surveys. In *Proceedings of the Survey Research Methods Section, American Statistical Association* (pp. 552-554). Alexandria, VA: American Statistical Association.

Eyerman, J., Bowman, K., Butler, D., & Wright, D. (2002). *The impact of incentives on cooperation and data collection costs: Results from the 2001 National Household Survey on Drug Abuse incentive experiment*. Paper presented at the annual conference of the American Association for Public Opinion Research, St. Pete Beach, FL.

Kennet, J. M., Gfroerer, J. Bowman, K. R., Martin, P. C., & Cunningham, D. (2003). *Effects of a $30 incentive on response rates and costs in the 2002 National Survey on Drug Use and Health*. Paper presented at the American Association for Public Opinion Research, Nashville, TN.

Kroutil, L., & Myers, L. (2002). Development of editing rules for CAI substance use data. In J. Gfroerer, J. Eyerman, & J. Chromy (Eds.), *Redesigning an ongoing national household survey: Methodological issues* (pp. 85-109, DHHS Publication No. SMA 03-3768). Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies. [Available as a PDF at http://www.oas.samhsa.gov/nhsda/methods.cfm#Reports]

Office of Applied Studies. (2001). *National Household Survey on Drug Abuse: 1999 nonresponse analysis report*. Rockville, MD: Substance Abuse and Mental Health Services Administration. [Available as a PDF at http://www.oas.samhsa.gov/nhsda/methods.cfm#Pre2k]

Office of Applied Studies. (2002, July). *2001 National Household Survey on Drug Abuse: Incentive experiment combined quarter 1 and quarter 2 analysis*. Rockville, MD: Substance Abuse and Mental Health Services Administration. [Available as a PDF at http://www.oas.samhsa.gov/nhsda/methods.cfm and http://www.oas.samhsa.gov/nhsda/methods/incentive.pdf]

Research Triangle Institute. (2001). *SUDAAN user's manual: Release 8.0*. Research Triangle Park, NC: Author.

RTI International. (2003). *2001 National Household Survey on Drug Abuse: Methodological resource book* (Report No. RTI 7190, prepared for the Substance Abuse and Mental Health Services Administration, Office of Applied Studies, under Contract No. 283-98-9008, Deliverable No. 28). Research Triangle Park, NC: Author. [Available at http://www.oas.samhsa.gov/nhsda/methods.cfm#2k1]

Wright, D., Bowman, K., Butler, D., & Eyerman, J. (2002). *Nonresponse bias from the National Household Survey on Drug Abuse incentive experiment*. Presented at the annual meeting of the American Association for Public Opinion Research, St. Pete Beach, FL.

4. Effects of the September 11, 2001, Terrorist Attacks on NSDUH Response Rates

Madeline E. McNeeley, Dawn Odom,[1] Julie Stivers, Peter Frechtel, Michael Langer, and Jim Brantley
RTI International

Dicy Painter and Joseph Gfroerer
Substance Abuse and Mental Health Services Administration

Introduction

The terrorist attacks on September 11, 2001, and the subsequent incidents of anthrax poisoning[2] have had a profound impact on the behaviors and attitudes of Americans (National Institute on Drug Abuse [NIDA], 2002; Office of Applied Studies [OAS], 2002; Schlenger et al., 2002; Silver, Holman, McIntosh, Poulin, & Gil-Rivas, 2002). Media reports following the attacks demonstrated such disparate effects as increasingly vocal support of government, some hostility toward Arab Americans and Muslims, and a heightened desire to be with family and friends. Travel within metropolitan New York City and Washington, DC, was restricted, mail from unfamiliar sources was considered potentially harmful, and people generally placed a greater emphasis on security issues. This new social environment may have had both positive and negative effects on interviewers' ability to make contact with respondents to the National Survey on Drug Use and Health (NSDUH)[3] and on those respondents' willingness to cooperate. Specifically, it was perceived that response rates may have been affected in two major ways, as shown in ***Figure 4.1***.

Because the survey is conducted year-round, has extensive call record data in addition to interview data, and has a sample large enough to allow conclusions to be drawn about small subpopulations, it is an ideal vehicle for studying the consequences of the terrorist attacks on survey research.

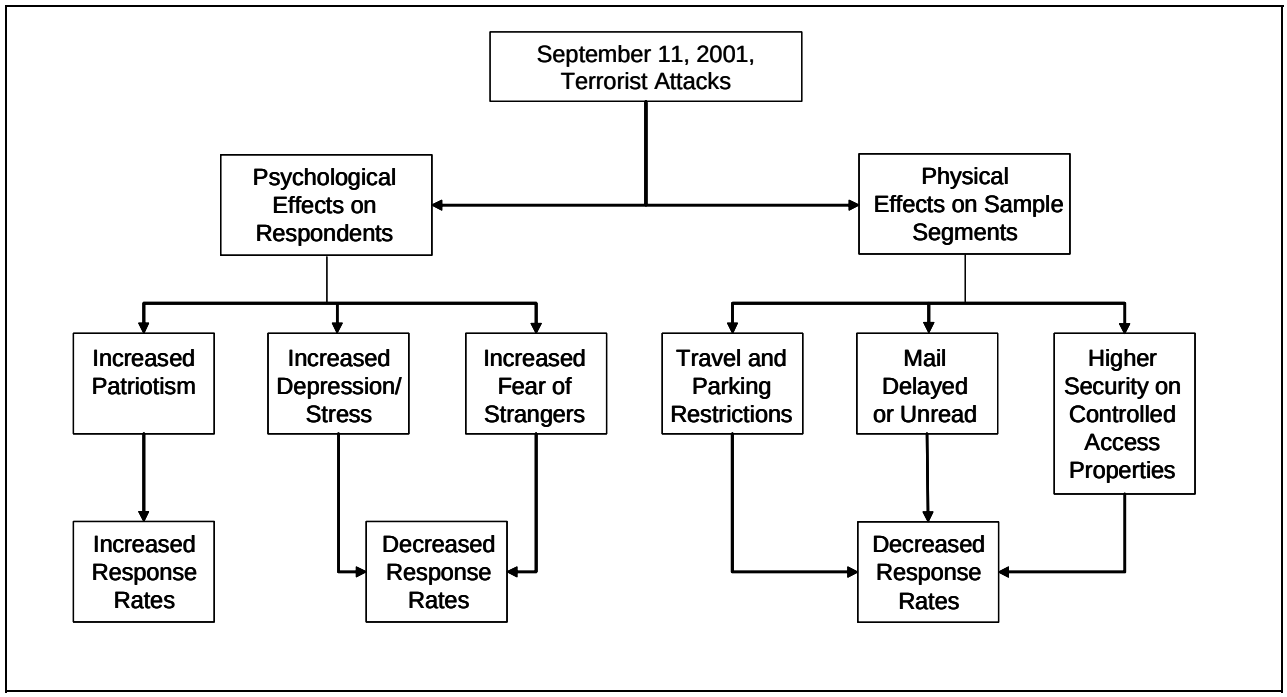
Several studies have examined the effects of national tragedies on individuals' psychological well-being. For example, Sheatsley and Feldman (1964) studied the public response to the assassination of President John F. Kennedy on November 22, 1963. These researchers investigated individual behaviors and emotional responses resulting from the assassination. They

[1] Now with Inveresk.

[2] The anthrax story first broke in September 2001 when an envelope containing anthrax spores was mailed to the offices of NBC-TV approximately 1 week after the September 11th terrorist attacks. Reports of anthrax poisoning continued through November 2001 (CNN.com, n.d.).

[3] Prior to 2002, the survey was called the National Household Survey on Drug Abuse (NHSDA). In this report, it is referred to as NSDUH, regardless of the survey year.

Figure 4.1 Perceived Influences of the September 11, 2001, Terrorist Attacks on NSDUH Response Rates



found that after Kennedy's assassination, measures of positive affect ("feelings that things were going their way" or "being proud of an accomplishment") were lower than at any previous time in the 1960s. After the Oklahoma City bombing on April 19, 1995, Smith, Christiansen, Vincent, and Hann (1999) conducted a study comparing stress and psychological distress of adults in the Oklahoma City metropolitan statistical area (MSA) and adults in a control area. The study found more stress, psychological distress, posttraumatic stress disorder (PTSD) components, and intrusive thoughts in the Oklahoma City MSA than in the control area.

In the 2 weeks following the September 11th attacks on the World Trade Center and the Pentagon, the National Opinion Research Center (NORC) conducted a study called the Public Response to a National Tragedy (Smith, Rasinski, & Toce, 2001). This was a telephone study involving a national sample and subsamples in New York City, Washington, DC, and Chicago. The study focused on individuals' behavior and communications, psychosomatic and affective responses, and political attitudes. Results were compared with data from both the post-Kennedy assassination study (Sheatsley & Feldman, 1964) and the General Social Survey (Smith & Jarkko, 1998). Preliminary findings indicated that after September 11th, there were increases in positive feelings, such as national pride and faith in human nature. Results for both positive and negative effects were compared across different demographic groups, and for the most part differences observed between demographic groups before the attacks remained after the attacks. The overall study had a response rate of 52 percent, with a 56 percent response rate for the national portion of the survey, 50 percent for New York, 41 percent for Washington, DC, and 51 percent for the Chicago area.

Schuster et al. (2001) reported on a national study conducted in the days following the September 11th attacks. This study investigated the emotional reactions of U.S. adults and their

perceptions of their children's emotional response to the events of September 11th. An estimated 44 percent of the adults surveyed reported at least one substantial symptom of stress; 90 percent had at least one symptom of stress to some degree. This computer-assisted telephone interviewing (CATI) study, conducted between September 14 and 16, 2001, obtained a 73 percent cooperation rate among known eligible households.

Schlenger et al. (2002) assessed psychological symptom levels in the United States following the events of September 11th using a Web-based, nationally representative survey with oversamples in the New York City and Washington, DC, metropolitan areas. The researchers found that 4.0 percent of people in the United States showed symptoms of probable PTSD, but that prevalence was much higher in the New York City area (11.2 percent). However, a broader measure of overall distress levels across the country was within expected ranges. Additionally, more than 60 percent of adults in New York City households with children reported that one or more children were upset by the attacks.

It is important to consider the perspective of potential survey respondents when measuring survey participation (Groves & Couper, 1998). The individuals whose psychological well-being and outlook are affected by an event like September 11th also are potential survey participants, so a national tragedy may potentially have an effect on survey respondents' willingness or availability to participate. The psychological effects of September 11th, as well as physical effects in the form of travel disruptions and heightened security measures, could have had real effects on the ability of survey staff to contact and gain participation from respondents.

The purpose of this study is to determine whether the events of September 11th had an effect on NSDUH screening response rates (SRRs) and interview response rates (IRRs). In addition to the preliminary statistical analysis of response rate data, logistic regression models were run to provide a more refined analysis. Field interviewer (FI) focus groups also were used to assess changes in the logistics of FI activity and in the use of the lead letter. To capture heightened concerns about security, FIs also were asked about increases in controlled access problems and changes in the mode of contact with screening and interview respondents.

Method

In Quarter 4 (Q4) of 2001, a New York City area sample supplement was implemented to increase the precision of prevalence estimates. The New York City area was identified as all of the New York City metropolitan areas in New York, Connecticut, and New Jersey (the New York City consolidated metropolitan statistical area [NYC CMSA]). The sample size was increased in these areas, increasing the person sample size from approximately 900 to 1,500. Due to the increased sample size in the New York City CMSA, traveling field interviewers (TFIs) were brought in to assist the local FIs in completing the additional caseload. The TFIs' overall response rates (ORRs) were considerably higher than those of the regular FI workforce (90.56 vs. 52.94 percent; see ***Table 4.1***). This difference in ORRs is expected because TFIs are specifically selected for their high levels of experience and proficiency. Cases assigned to the TFIs made up a relatively small portion of the full sample, so they increased the ORR for the area by only 1.6 percentage points to 54.54 percent.

Four geographic areas[4] were identified for study of the September 11th effect:

1. Nation;
2. New York City CMSA;
3. Washington, DC, primary metropolitan statistical area (DC PMSA); and
4. Nation excluding the New York City CMSA and Washington, DC, PMSA.

Table 4.1 Weighted Response Rates, by Field Interviewer Type, New York City CMSA Only, Quarter 4 2001 NSDUH

FI Type	Screening		Interviewing		Overall Rate (%)
	N	Rate (%)	*N*	Rate (%)	
TFIs	211	95.20	78	95.13	90.56
Non-TFIs	4,093	81.95	1,985	64.60	52.94
All FIs	4,304	82.63	2,063	66.00	54.54

FI = field interviewer; TFI = traveling field interviewer.

The research plan applied two methodologies: a series of focus groups with FIs and a comparative analysis of response rates. Each of these methodologies is discussed in further detail below.

Focus Groups

Four focus groups were held with a sample of FIs to examine nonresponse and field activity issues related to September 11th. Each focus group included five or six interviewers. Field supervisors (FSs) from the New York City CMSA, the Washington, DC, PMSA, and other areas across the country were asked to help generate a list of FIs who had worked a considerable number of cases both before and after September 11, 2001. This was done to ensure that the participants would be able to contribute to a discussion of how their work changed after the attacks. During these conversations with FSs from the New York City CMSA and the Washington, DC, PMSA, they also were asked informally to describe the effects they perceived that September 11th had on their FIs' work patterns and response rates.

Project staff called FIs from these lists, secured their agreement to participate, and gave the participants their assigned conference line phone numbers and call-in times. ***Table 4.2*** summarizes the composition of each focus group.

Table 4.2 Composition of Field Interviewer Focus Groups

Date	Number of Participants	Geographic Areas Represented
3/13/02	5 FIs and 1 TFI	Illinois, Michigan (1 FI and 1 TFI), Pennsylvania, Texas, and Utah
3/19/02	5 FIs	Southern Maryland and Washington, DC
3/20/02	5 FIs	Alabama, California, Illinois, Indiana, and Montana
3/21/02	5 FIs	Greater New York City area, including Connecticut, New Jersey, and New York

FI = field interviewer; TFI = traveling field interviewer.

[4] Initially, the study plan included separate analyses for Manhattan and for New York City (five boroughs only), but sample sizes for these areas were too small to permit valid comparisons.

Response Rate Comparison

With the increased New York City area sample size in Q4 2001 and the annual administration of the survey, it was possible to compare response rates prior to and subsequent to the events of September 11th. Several limiting factors drove the design of the analysis. First, September 11th occurred in the third month of Q3 2001, when the national field staff was in "cleanup" mode and most of the screening and interviewing work for Q3 had already been completed, so one could not expect to find noticeable effects until Q4. (This quarterly replicate design might have limited the ability to detect change in that it places the measure of change somewhat distant from the hypothesized cause.) One also could not determine whether the effects of September 11th would have persisted into 2002 data collection because survey design changes that could affect response rates were implemented in 2002, such as a $30 incentive payment and the change in the name of the survey.[5] Therefore, the analysis was focused on Q4 2001 response rates (the post-September 11th data collection period). Two comparisons using Q4 as the basis to distinguish any effects of September 11th were considered:

- $Q_{4,2001} - Q_{1\text{-}3,2001}$ and
- $(Q_4 - Q_{1\text{-}3})_{2001} - (Q_4 - Q_{1\text{-}3})_{2000}$.

However, an incentive experiment was conducted in a subset of Q1 2001 and Q2 2001 that it was felt would contaminate these comparisons. To compensate for this limitation and also for seasonality effects, it was decided to analyze the September 11th effect by only comparing Q4 2001 with Q4 2000 using the equation $Q4_{2001}$ - $Q4_{2000}$.

Two distinct methods were used for statistically testing the September 11th effect on response rates. Both methods used weights based on the design of the survey, and both used SAS®-callable SUDAAN®, a software package that conveniently allows the user to account for the full multistage design of a survey.[6] The first was a *t* test that directly compared the response rates between Q4 2000 and Q4 2001 without controlling for any other variables. The results of these tests are displayed in ***Tables 4.3*** and ***4.4***. These tests are called "difference tests" in the following text. The second statistical method was logistic regression. The response variables in the logistic regression models were indicators of whether the subject responded to the screening or the interview. Each model included a "September 11th" variable as a covariate, as well as several other covariates. The "September 11th" variable had two levels: one for before September 11th (i.e., Q4 2000) and one for after September 11th (i.e., Q4 2001). Examining the significance of the parameter associated with the "September 11th" variable in the model allowed a statistical test for a change in response rates while controlling for all the other covariates in the models.

The other covariates used in the logistic regression models of the SRR included a number of segment-level variables: concentrations of Hispanic, non-Hispanic black, and owner-occupied dwelling units, as well as census region and population density for the "Nation" and "Nation Excluding NYC CMSA and DC PMSA" models. Also included were measures of FI age,

[5] Chapter 2 of this volume provides a detailed description of the changes implemented in the 2002 survey.

[6] Details about the statistical methods can be found in the *SUDAAN® User's Manual: Release 8.0* (Research Triangle Institute [RTI], 2001). Additional references are provided in this user's manual. SAS® software is a registered trade mark of SAS Institute, Inc., and SUDAAN® is a registered trade mark of RTI.

Table 4.3 Weighted Screening Response Rates, by Region, 2000-2001 NSDUH

Region	Quarter 4 2000		Quarter 4 2001		Difference	
	N	Rate (%)	*N*	Rate (%)	(%)	*P* Value
Nation	39,532	93.10	44,116	91.52	-1.59	<0.001
Nation Excluding NYC CMSA and DC PMSA	35,587	93.44	38,524	92.25	-1.19	<0.001
NYC CMSA only	2,206	88.74	4,304	82.63	-6.11	<0.001
DC PMSA only	1,739	92.35	1,288	87.01	-5.34	0.048

NYC CMSA = New York City consolidated metropolitan statistical area.
DC PMSA = Washington, DC, primary metropolitan statistical area.

Table 4.4 Weighted Interview Response Rates, by Region, 2000-2001 NSDUH

Region	Quarter 4 2000		Quarter 4 2001		Difference	
	N	Rate (%)	*N*	Rate (%)	(%)	*P* Value
Nation	19,428	73.82	23,598	71.97	-1.85	0.036
Nation Excluding NYC CMSA and DC PMSA	18,058	73.64	21,120	72.30	-1.34	0.142
NYC CMSA only	837	72.65	2,063	66.00	-6.65	0.110
DC PMSA only	533	84.66	415	77.71	-6.95	0.137

NYC CMSA = New York City consolidated metropolitan statistical area.
DC PMSA = Washington, DC, primary metropolitan statistical area.

race/ethnicity, gender, and number of screenings completed (as a proxy for FI experience),[7] and whether the FI was a TFI. Models of the IRR controlled on segment-level variables (census region and population density), respondent-level (race/ethnicity, age, and gender), and interviewer-level variables (age, race/ethnicity, gender, number of completed screenings, and whether the FI was a TFI).

Field Interviewer Focus Groups

FIs in each focus group were asked to address a variety of issues related to September 11th, including changes in controlled access issues, mode of contact with the respondents (i.e., speaking to respondents through closed doors, screen doors, intercoms), respondents' use of the lead letter, and logistical issues, such as parking and traffic. This section presents the findings of the focus groups. Some topics of discussion are not specifically mentioned below, indicating that the FIs saw no effect of September 11th on those issues.

The FIs from outside New York City and Washington, DC, indicated that the events of September 11th temporarily increased respondents' willingness to participate, particularly among male respondents. Other than that small difference, they saw no other real effects of September 11th.

FIs are generally encouraged to complete their assignments by the end of the second month of each quarter in order to leave sufficient time for "cleanup" of the current quarter's caseload

[7] Number of screenings was cumulative within each year only, not summed across years.

and preparation for the next quarter. Historically, NSDUH has had lower response rates in New York City than in much of the rest of the Nation, and by September 11th their third quarter fieldwork was not yet complete. On September 11th, 12th, and 13th, no New York City area FIs went into the field; starting on September 14th, limited fieldwork began. This reduced the number of days FIs had available to them to finish their Q3 caseload. The FIs said that, because of the logistical problems they continued to encounter, they did not return to normal fieldwork until several weeks after September 11th. Because Q4 data collection began on October 1st, the events of September 11th also had a direct impact on the amount of time available for and the logistical challenges to completing the Q4 caseload.

The FIs from the New York City CMSA focus groups also indicated that September 11th had a temporary but positive effect on respondents' willingness to participate, once the FIs had been able to make contact with the households and identify the respondents. These FIs stated that they did face some additional challenges as a direct result of the events of September 11th, such as increased transportation problems immediately after the attacks. One FI thought that readership of the lead letter was greater after September 11th due to the government logo on the envelope and letterhead, but most FIs thought that more respondents were suspicious of the letter and threw it away without opening it.

Field management staff noted that once interviewing resumed, interviewers reported that finding people at home was a major problem, particularly in apartments and other single-person households. A common observation made by the New York interviewing staff was that single people wanted to be with friends, and this often would occur somewhere other than the home. They felt that controlled or restricted access problems increased, not due to more security by doormen or management, but as a result of a reluctance of individual tenants to open their doors or answer their intercoms.

In contrast with New York City, Washington, DC, has a history of high NSDUH response rates. Most of its Q3 fieldwork was completed before September 11th. As a result, field management staff felt that any effect of the terrorist attacks on response rates was not seen until Q4 2001. Similar to interviewers in the New York City CMSA, interviewers in Washington, DC, and southern Maryland did not work on September 11th, 12th, and 13th. On September 15th, fieldwork began at a relatively normal rate. Management cited that increased traffic, parking restrictions, closed roads, and problems with public transportation had a negative impact on productivity after September 11th that lasted through November 2001.

The FIs from the Washington, DC, PMSA focus groups reported that their ability to contact certain households was somewhat reduced by heightened security on college campuses. In addition, many military reservists were called up to active duty after September 11th, making them ineligible for the survey and potentially affecting the composition of the sample in the DC area.

Other than the observations made above, the FIs noticed no real differences in their fieldwork after September 11th. Specifically, the FIs did not notice much change in the mode of contact with the respondents, and most FIs believed that households' treatment of the lead letter was not affected by the anthrax threat and investigation. The FIs found that for the most part, screening respondents seemed to open the lead letter at about the same rate as before the cases of anthrax

poisoning occurred. A possible explanation for this is that the lead letter is delivered in an envelope bearing the logo of the U.S. Department of Health and Human Services, which may have increased household members' confidence that the letter was safe.

Results

Analyses of the effects of the terrorist attacks on NSDUH response rates included descriptive analyses of differences in SRRs and IRRs between Q4 2000 and Q1 2001, patterns of nonresponse, and modeling of response rates.

Descriptive Analysis

As described earlier, the effect of September 11th on response rates was calculated by comparing Q4 response rates from 2000 with Q4 response rates from 2001 ($Q4_{2001}$ - $Q4_{2000}$). This September 11th comparison is examined here through descriptive analysis and significance testing where applicable.

The preliminary analysis consisted of screening, interview, and overall response rate comparisons (SRR, IRR, and ORR, respectively). Sample weights were used in the analysis in order to account for the complex survey design.[8] Response rates were produced for various demographics where applicable, including census region, population density, and respondent race, age, and gender. ***Tables 4.5*** and ***4.6*** summarize the statistically significant differences in SRR and IRR[9] for each area; the full results are presented in ***Tables 4.7*** to ***4.10***.

Table 4.5 Summary of Significant Differences in Weighted Screening Response Rates, by Region, 2000-2001 NSDUH

Region	Subgroup	Difference	*P* Value
Nation	Full sample	-1.59	<0.001
	Northeast	-3.31	<0.001
	Midwest	-1.17	0.040
	South	-2.10	<0.001
	≥1,000,000 population	-2.21	<0.001
	50K-999,999 population	-1.37	0.003
Nation Excluding NYC CMSA and DC PMSA	Full sample	-1.19	<0.001
	Northeast	-1.82	0.006
	Midwest	-1.17	0.040
	South	-1.91	<0.001
	≥1,000,000 population	-1.28	0.020
	50K-999,999 population	-1.43	0.002
NYC CMSA only	Full sample	-6.11	<0.001
DC PMSA only	Full sample	-5.34	0.048

NYC CMSA = New York City consolidated metropolitan statistical area.
DC PMSA = Washington, DC, primary metropolitan statistical area.

[8] For screening rates, dwelling unit (DU)–level design weight components were used. For interview rates, in addition to DU-level design weight components, DU–level nonresponse, poststratification, and extreme weight treatment components and person-level design weight components were included.

[9] Significance testing was conducted only for SRRs and IRRs.

Table 4.6 Summary of Significant Differences in Weighted Interview Response Rates, by Region, 2000-2001 NSDUH

Region	Subgroup	Difference	*P* Value
Nation	Full sample	-1.85	0.036
	Northeast	-4.31	0.033
	Black, non-Hispanic	-4.74	0.050
	12 to 17 years old	-2.27	0.014
	18 to 25 years old	-4.10	<0.001
Nation Excluding NYC CMSA and DC PMSA	18 to 25 years old	-3.44	0.003
DC PMSA only	White, non-Hispanic	-11.53	0.044
	18 to 25 years old	-18.38	<0.001
	35 to 49 years old	-15.51	0.018

NYC CMSA = New York City consolidated metropolitan statistical area.
DC PMSA = Washington, DC, primary metropolitan statistical area.

In addition to the general response rate analysis, other aspects of nonresponse for the four areas of interest were investigated using the same Q4 2000 versus Q4 2001 approach. Contact rates were examined in order to establish whether these rates changed after September 11th. Also, it was hypothesized that the at-home patterns of respondents would change in the New York City and Washington, DC, areas. To investigate this, the mean number of attempts required to make contact with the screening respondent was calculated.

Preliminary Response Rate Analysis

Nation. Decreases were expected in SRRs and IRRs for the full national sample between Q4 2000 and Q4 2001. The Northeast (which includes the New York City CMSA) and the South (including the Washington, DC, PMSA) were the two census regions most directly affected by the terrorist attacks; therefore, the greatest reductions were expected in these areas, as well as in large metropolitan areas. ***Table 4.7*** presents the national weighted response rates (SRR, IRR, and ORR) by the various geographic and demographic variables. There was a small but significant national decline in SRRs (d = -1.59) and IRRs (d = -1.85) between Q4 2000 and Q4 2001. The biggest differences in the ORR were seen in the Northeast census region, which had a difference of 6.19 percentage points between the two quarters, and the South, with a difference of 3.74 percentage points. By comparison, the Midwest and West showed ORR differences of only 0.50 and 0.81 percentage points, respectively. In the SRR, however, changes in the Northeast, Midwest, and South were all significant, though the difference in the Midwest was smaller than in the other two regions (d_{NE} = -3.31, d_{MW} = -1.17, d_S = -2.10). The Northeast was the only census region to see a significant change in IRR (d = -4.31).

Metropolitan areas showed a reduction in ORR of about 3 percentage points from Q4 2000 to Q4 2001. Their differences in SRR were significant (d = -2.21 for MSAs with populations of 1,000,000 or more, d = -1.37 for other MSAs), but their differences in IRR were not significant. Non-MSAs showed no significant differences. Again, these results were anticipated as the attacks occurred in the metropolitan areas of New York City and Washington, DC.

IRRs also were computed for respondent demographics (race/ethnicity, age, and gender). Race/ethnicity was defined as white (non-Hispanic), black (non-Hispanic), Hispanic, or "other," where "other" includes all non-Hispanic persons who did not define themselves as white or black

Table 4.7 National Weighted Response Rates, 2000-2001 NSDUH

Response Rates	Quarter 4 2000		Quarter 4 2001		Difference	
	N	Rate (%)	*N*	Rate (%)	(%)	*P* Value
Screening Response Rates (SRRs)	**39,532**	**93.10**	**44,116**	**91.52**	**-1.59**	**<0.001**
Region						
Northeast	9,588	91.18	11,465	87.87	-3.31	<0.001
Midwest	10,750	93.05	11,843	91.88	-1.17	0.040
South	11,631	93.99	12,599	91.90	-2.10	<0.001
West	7,563	93.47	8,209	93.73	0.27	0.670
Population Density						
≥1,000,000 population	15,256	91.93	16,835	89.72	-2.21	<0.001
50K-999,999 population	13,386	93.25	15,527	91.88	-1.37	0.003
Non-MSA	10,890	95.13	11,754	94.40	-0.73	0.067
Interview Response Rates (IRRs)	**19,428**	**73.82**	**23,598**	**71.97**	**-1.85**	**0.036**
Region						
Northeast	4,496	72.75	5,628	68.44	-4.31	0.033
Midwest	5,612	72.47	6,614	72.85	0.37	0.777
South	5,469	75.62	6,569	73.27	-2.35	0.137
West	3,851	73.05	4,787	71.98	-1.07	0.601
Population Density						
≥1,000,000 population	6,993	72.23	8,514	70.29	-1.94	0.188
50K-999,999 population	6,946	73.97	8,767	71.67	-2.30	0.128
Non-MSA	5,489	76.63	6,317	75.59	-1.04	0.524
Respondent Race/Ethnicity						
White, non-Hispanic	13,967	73.37	16,447	71.70	-1.67	0.108
Black, non-Hispanic	2,348	76.87	2,795	72.14	-4.74	0.050
Hispanic	2,146	77.97	3,028	76.68	-1.29	0.610
Other	967	64.47	1,328	65.04	0.56	0.909
Respondent Age						
12-17	6,710	83.37	7,202	81.10	-2.27	0.014
18-25	6,327	77.43	8,137	73.33	-4.10	<0.001
26-34	2,541	73.91	2,365	74.07	0.16	0.913
35-49	1,889	74.11	3,573	71.64	-2.47	0.121
≥50	1,961	69.43	2,321	68.18	-1.25	0.525
Respondent Gender						
Female	9,860	74.38	12,026	72.94	-1.44	0.246
Male	9,568	73.21	11,572	70.89	-2.32	0.052
Overall Response Rates (ORRs)	**Rate (%)**		**Rate (%)**		**Difference[1] (%)**	
	68.73		**65.86**		**-2.87**	
Region						
Northeast	66.33		60.14		-6.19	
Midwest	67.43		66.93		-0.50	
South	71.07		67.33		-3.74	
West	68.28		67.47		-0.81	
Population Density						
≥1,000,000 population	66.40		63.06		-3.34	
50K-999,999 population	68.98		65.85		-3.13	
Non-MSA	72.89		71.36		-1.53	

MSA = metropolitan statistical area.
Note: Respondent age, race/ethnicity, and gender demographics were not available for screening and overall response rates.
[1]Differences in ORR were not tested for statistical significance.

(i.e., Asian, Pacific Islander, etc.). Nationally, black respondents were the only racial/ethnic group to show a significant decrease in IRR from 76.87 percent in Q4 2000 to 72.14 percent in Q4 2001. Age was categorized into five levels: 12 to 17, 18 to 25, 26 to 34, 35 to 49, and 50 or older. The most notable and significant difference in IRR was among 18 to 25 year olds, with a reduction of 4.10 percentage points. One possible explanation for this is that many individuals in this age group attend college, where tightened security may have hampered FIs. The difference among the 12 to 17 group also was significant, though not as large (d = -2.27). There were no significant differences by gender of respondent.

New York City CMSA. ***Table 4.8*** displays the New York City CMSA's weighted response rates for Q4 2000 and Q4 2001. It was assumed that survey respondents in the New York City CMSA generally would have been home less often and have been less willing to participate in the survey due to the events of September 11th, specifically because of the proximity of the attacks on the World Trade Center. In fact, there was a significant decrease in SRR (d = -6.11). The decrease in IRR, though larger than that in the national analysis, was not significant due to the smaller sample size. On the whole, this area's ORR decreased nearly 10 percentage points from 64.47 to 54.54 percent. Possible explanations for this major response rate decrease might include an increase in "not at homes" (see the Categories of Nonresponse section later in this chapter), adverse effects on the psychological well-being of New York City residents (NIDA, 2002; Schlenger et al., 2002; Silver et al., 2002), and the fact that many of the sampled dwelling units were difficult or even impossible to reach due to the destruction.

Table 4.8 New York City CMSA Weighted Response Rates, 2000-2001 NSDUH

Response Rates	Quarter 4 2000		Quarter 4 2001		Difference	
	N	Rate (%)	*N*	Rate (%)	(%)	*P* Value
Screening Response Rates (SRRs)	**2,206**	**88.74**	**4,304**	**82.63**	**-6.11**	**<0.001**
Interview Response Rates (IRRs)	**837**	**72.65**	**2,063**	**66.00**	**-6.65**	**0.110**
Respondent Race/Ethnicity						
White, non-Hispanic	487	71.37	1,140	64.57	-6.80	0.166
Black, non-Hispanic	139	83.10	297	67.90	-15.19	0.057
Hispanic	148	72.04	482	73.91	1.86	0.868
Other	63	62.45	144	55.94	-6.50	0.572
Respondent Age						
12-17	292	84.15	641	78.14	-6.01	0.087
18-25	225	73.89	749	64.69	-9.20	0.054
26-34	140	77.24	220	71.93	-5.31	0.278
35-49	78	70.02	299	68.16	-1.86	0.787
≥50	102	69.78	154	57.69	-12.09	0.154
Respondent Gender						
Female	415	72.05	1,060	66.90	-5.15	0.358
Male	422	73.24	1,003	64.86	-8.38	0.072
Overall Response Rates (ORRs)	**Rate (%) 64.47**		**Rate (%) 54.54**		**Difference[1] (%) -9.93**	

CMSA = consolidated metropolitan statistical area.
Note: Respondent age, race/ethnicity, and gender demographics were not available for screening and overall response rates.
[1]Differences in ORR were not tested for statistical significance.

Upon closer examination, several subgroups were identified that showed large IRR differences. As with the main IRR figure, these differences were generally larger than those seen nationally. However, because of the smaller sample size in the New York City CMSA, none of these differences was significant.

Washington, DC, PMSA. The Washington, DC, PMSA also was directly affected on September 11th due to the attack on the Pentagon, so effects similar to those on the New York City CMSA were expected. ***Table 4.9*** displays this area's weighted response rates for Q4 2000 and Q4 2001. The ORR in the Washington, DC, PMSA decreased 10.57 percentage points from 78.18 percent in Q4 2000 to 67.61 percent in Q4 2001. Similar decreases were seen for the individual screening and interview rates. Similar to the patterns in the New York City CMSA, there was a significant decrease in SRR for the Washington, DC, PMSA (d = -5.34), and a large decrease in IRR that was not significant due to small sample size (d = -6.95).

Table 4.9 Washington, DC, PMSA Weighted Response Rates, 2000-2001 NSDUH

Response Rates	Quarter 4 2000		Quarter 4 2001		Difference	
	N	Rate (%)	*N*	Rate (%)	%	*P* Value
Screening Response Rates (SRRs)	**1,739**	**92.35**	**1,288**	**87.01**	**-5.34**	**0.048**
Interview Response Rates (IRRs)	**533**	**84.67**	**415**	**77.71**	**-6.95**	**0.137**
Respondent Race/Ethnicity						
White, non-Hispanic	201	87.66	169	76.12	-11.53	0.044
Black, non-Hispanic	245	87.52	166	77.88	-9.63	0.194
Hispanic	53	95.57	42	91.94	-3.63	0.528
Other	34	61.96	38	81.09	19.13	0.463
Respondent Age						
12-17	228	89.30	139	77.50	-11.80	0.127
18-25	150	92.96	128	74.58	-18.38	<0.001
26-34	48	67.50	49	78.54	11.04	0.346
35-49	55	94.59	65	79.08	-15.51	0.018
≥50	52	75.23	34	76.48	1.26	0.932
Respondent Gender						
Female	279	87.96	216	77.88	-10.09	0.056
Male	254	81.60	199	77.57	-4.03	0.485
Overall Response Rates (ORRs)	**Rate (%)**		**Rate (%)**		**Difference[1] (%)**	
	78.18		**67.61**		**-10.57**	

PMSA = primary metropolitan statistical area.
Note: Respondent age, race/ethnicity, and gender demographics were not available for screening and overall response rates
[1]Differences in ORR were not tested for statistical significance.

Unlike in the national and New York City CMSA areas, the Washington, DC, PMSA's showed a significant change in IRR among white respondents (d = -11.53). Differences by gender were not significant. Among 35 to 49 year olds, there was a significant decrease in IRR (d = -15.51). There also was a significant decrease in IRR among the 18 to 25 year olds (d = -18.38). Again, this could have been caused by the behaviors of that age group, such as spending less time at home. The most likely reason in this case, however, is that many 18 to 25 year olds may live in apartment complexes and on college campuses, where tighter security could prohibit

interviewers from successfully gaining entrance. FIs in the Washington, DC, PMSA encountered an unusually large number of controlled access problems during Q4 2001.

Nation Excluding New York City CMSA and Washington, DC, PMSA. Some dramatic differences were evident in the Washington, DC, and New York City areas, but it was not clear whether those areas alone had been responsible for the results seen in the national analysis. For this reason, the national rates excluding the New York City CMSA and Washington, DC, PMSA were reanalyzed. In this case, it was expected that the decline in response rates would remain significant, but that the magnitude of the difference would not be as great as when the New York City CMSA and Washington, DC, PMSA were included. ***Table 4.10*** displays the weighted response rates by various demographics for this analysis.

The ORR in Q4 2000 was 68.81 percent, dropping 2.11 percentage points to 66.70 percent in Q4 2001. This is a somewhat smaller reduction than the overall national estimate. The difference in SRR, though smaller than in the full national sample, was still significant (d = -1.19); the difference in IRR, however, was not significant. Parallel to the full national estimates, the Northeast and South demonstrated the biggest ORR decreases (4.06 and 3.27 percentage points, respectively). The changes in SRR for the Northeast, Midwest, and South were all significant (d_{NE} = -1.82, d_{MW} = -1.17, d_S = -1.91), though the differences in the Northeast and South were not as large as they had been before the exclusion of New York City and Washington, DC. There were no significant differences in IRR by census region.

After excluding the Washington, DC, PMSA and New York City CMSA, the decreases in SRR remained significant for the highest density populations (50,000 or more). The IRRs by respondent race and gender showed no significant difference between Q4 2000 and Q4 2001. The only significant difference by subgroup was among 18 to 25 year olds, who saw a decrease of 3.44 percentage points. This is a somewhat smaller difference than the one seen in the full national estimates.

Additional Analyses

At-Home Patterns of Respondents. At-home patterns were investigated by looking at the mean number of attempts required to make first contact with screening respondents (SRs) who were eventually contacted for the study. ***Table 4.11*** illustrates the results of this investigation. (Note that this differs from other analyses, such as the categories of nonresponse presented in ***Tables 4.12*** and ***4.13***, in that ***Table 4.11*** is unweighted and only includes calls leading up to the initial contact with the SR.) In the New York City CMSA, there was a small but significant decrease in the mean number of attempts required to make initial contact with an appropriate SR. Although a smaller percentage of SRs were contacted after September 11th in the New York City CMSA (as seen in the next analysis), the SRs who were eventually contacted were easier to locate. Otherwise, there were no significant results in this analysis.

Table 4.10 National Weighted Response Rates Excluding the New York City CMSA and the Washington, DC, PMSA, 2000-2001 NSDUH

Response Rates	Quarter 4 2000		Quarter 4 2001		Difference	
	N	Rate (%)	*N*	Rate (%)	%	*P* Value
Screening Response Rates (SRRs)	**35,587**	**93.44**	**38,524**	**92.25**	**-1.19**	**<0.001**
Region						
Northeast	7,382	92.50	7,161	90.68	-1.82	0.006
Midwest	10,750	93.05	11,843	91.88	-1.17	0.040
South	9,892	94.08	11,311	92.16	-1.91	<0.001
West	7,563	93.47	8,209	93.73	0.27	0.670
Population Density						
≥1,000,000 population	11,962	92.47	12,193	91.19	-1.28	0.020
50K-999,999 population	12,762	93.37	14,596	91.94	-1.43	0.002
Non-MSA	10,863	95.13	11,735	94.42	-0.71	0.075
Interview Response Rates (IRRs)	**18,058**	**73.64**	**21,120**	**72.30**	**-1.34**	**0.142**
Region						
Northeast	3,659	72.80	3,565	69.78	-3.03	0.156
Midwest	5,612	72.47	6,614	72.85	0.37	0.777
South	4,936	75.05	6,154	73.06	-1.99	0.227
West	3,851	73.05	4,787	71.98	-1.07	0.601
Population Density						
≥1,000,000 population	5,885	71.27	6,574	70.60	-0.67	0.670
50K-999,999 population	6,704	74.23	8,242	71.90	-2.33	0.138
Non-MSA	5,469	76.63	6,304	75.63	-1.00	0.542
Respondent Race/Ethnicity						
White, non-Hispanic	13,279	73.17	15,138	72.09	-1.08	0.316
Black, non-Hispanic	1,964	75.97	2,332	72.27	-3.69	0.154
Hispanic	1,945	78.59	2,504	76.86	-1.73	0.490
Other	870	64.82	1,146	65.65	0.83	0.874
Respondent Age						
12-17	6,190	83.21	6,422	81.37	-1.83	0.054
18-25	5,952	77.35	7,260	73.91	-3.44	0.003
26-34	2,353	73.76	2,096	74.12	0.36	0.819
35-49	1,756	73.67	3,209	71.75	-1.92	0.243
≥50	1,807	69.27	2,133	68.77	-0.50	0.806
Respondent Gender						
Female	9,166	74.25	10,750	73.33	-0.91	0.480
Male	8,892	72.99	10,370	71.17	-1.82	0.146
Overall Response Rates (ORRs)	**Rate (%)**		**Rate (%)**		**Difference[1] (%)**	
	68.81		**66.70**		**-2.11**	
Region						
Northeast	67.34		63.28		-4.06	
Midwest	67.43		66.93		-0.50	
South	70.60		67.33		-3.27	
West	68.28		67.47		-0.81	
Population Density						
≥1,000,000 population	65.90		64.38		-1.52	
50K-999,999 population	69.31		66.10		-3.21	
Non-MSA	72.90		71.41		-1.49	

CMSA = consolidated metropolitan statistical area; MSA = metropolitan statistical area; PMSA = primary metropolitan statistical area.

Note: Respondent age, race/ethnicity, and gender demographics were not available for screening and overall response rates.

[1] Differences in ORR were not tested for statistical significance.

Table 4.11 Unweighted Mean Number of Attempts to Contact Screening Respondents, 2000-2001 NSDUH

Region	Quarter 4 2000	Quarter 4 2001	Difference	*P* Value
National	3.09	3.04	-0.05	0.185
National Excluding NYC CMSA and DC PMSA	3.05	3.01	-0.04	0.225
NYC CMSA	3.53	3.10	-0.43	0.011
DC PMSA	3.40	4.02	0.62	0.088

NYC CMSA = New York City consolidated metropolitan statistical area.
DC PMSA = Washington, DC, primary metropolitan statistical area.
Note: Based on all visits made to eligible households, up to and including the first contact with an appropriate screening respondent.

Categories of Nonresponse. Nonresponse categories were examined by comparing the percentages of finalized screening and interview cases that had received a final code of refusal or "not-at-home" in Q4 2000 or Q4 2001 (see ***Tables 4.12*** and ***4.13***). Based on the FI focus groups, we expected screening and interviewing "not-at-home" codes to increase in the New York City CMSA and Washington, DC, PMSA, but did not expect to see a significant change nationwide. Fewer screening and interviewing refusals were anticipated due to a hypothesized increase in patriotic attitudes after September 11th.

Nationally, the percentages of screening respondents who were not at home differed only slightly ($d = 0.33$ percentage points), but this difference is statistically significant. There also was a significant, much larger increase in screening not-at-home codes in the New York City CMSA ($d = 2.62$). No significant differences were observed in interview not-at-home codes. Screening refusals increased in all four areas: national ($d = 1.03$), New York City CMSA ($d = 2.42$), Washington, DC, PMSA ($d = 2.49$), and national excluding these two areas ($d = 0.91$). Increases in interview refusals were significant nationally, both including ($d = 1.95$) and excluding ($d = 1.65$) New York City and Washington, DC.

The significance of the minor national increase in screening not-at-home codes can probably be explained by the large sample size; similarly, the small size of the Washington, DC, PMSA sample could explain why the observed differences were not statistically significant. The New York City CMSA screening not-at-home codes were as expected. The expected increases in interview not-at-home codes did not occur; this is probably due to interviewers' ability to make appointments and get "on the spot" interviews with selected respondents.

A decrease in refusal rates was anticipated in every area, which did not occur. This seems to indicate that the apparent increase in patriotic attitudes after September 11th did not necessarily translate into increased cooperation rates with government-sponsored surveys.

The data described in this section generally correspond with the anecdotal field evidence. Response rates in the New York City CMSA and Washington, DC, PMSA did decline as expected. Furthermore, the analysis shows that one of the most affected groups was the 18 to 25 year old group, which has a large number of single persons living alone. Both results agree with the field evidence cited above.

Table 4.12 Weighted Percentages of Screening Refusals and "Not-at-Homes," 2000-2001 NSDUH

Region	Percent of Refusals				Percent of "Not-at-Homes"			
	Q4 2000	Q4 2001	Difference	*P* Value	Q4 2000	Q4 2001	Difference	*P* Value
Nation	4.04	5.07	1.03	<0.001	2.09	2.42	0.33	0.028
Nation Excluding NYC CMSA and DC PMSA	3.83	4.74	0.91	<0.001	1.99	2.13	0.14	0.309
NYC CMSA	6.63	9.05	2.42	0.008	3.35	5.97	2.62	0.021
DC PMSA	4.69	7.18	2.49	0.044	2.36	4.07	1.71	0.285

NYC CMSA = New York City consolidated metropolitan statistical area.
DC PMSA = Washington, DC, primary metropolitan statistical area.
Note: Based on the final visit code made to eligible households.

Table 4.13 Weighted Percentages of Interview Refusals and "Not-at-Homes," 2000-2001 NSDUH

Region	Percent of Refusals				Percent of "Not-at-Homes"			
	Q4 2000	Q4 2001	Difference	*P* Value	Q4 2000	Q4 2001	Difference	*P* Value
Nation	15.60	17.55	1.95	0.005	5.76	5.74	-0.02	0.961
Nation Excluding NYC CMSA and DC PMSA	15.99	17.64	1.65	0.024	5.56	5.35	-0.21	0.626
NYC CMSA	13.23	18.20	4.97	0.080	8.61	10.34	1.73	0.507
DC PMSA	6.46	9.54	3.08	0.497	4.99	8.31	3.32	0.300

NYC CMSA = New York City consolidated metropolitan statistical area.
DC PMSA = Washington, DC, primary metropolitan statistical area.
Note: Based on the final visit code made to eligible persons.

Response Rate Modeling

This section presents the results of the logistic models described in previous sections. The parameter associated with the "September 11th" variable in each model allows for a statistical test of change in response rate from Q4 2000 to Q4 2001, while controlling on various segment-level, interviewer-level, and respondent-level variables. Specifically, the models of SRR included the following variables:

- concentration of non-Hispanic black residents in segment,
- concentration of owner-occupied dwelling units (DUs) in segment,
- concentration of Hispanic residents in segment,
- population density ("Nation" and "Nation Excluding New York City and Washington, DC," only),
- census region ("Nation" and "Nation Excluding New York City and Washington, DC," only),

- number of screenings completed by FI,
- age of FI,[10]
- gender of FI,
- race/ethnicity of FI[11] (all models except "Washington, DC, PMSA"), and
- travel status of FI.

The models of IRR included the following:

- race/ethnicity of respondent,
- age of respondent,
- gender of respondent,
- population density ("Nation" and "Nation Excluding New York City and Washington, DC," only),
- census region ("Nation" and "Nation Excluding New York City and Washington, DC," only),
- number of screenings completed by FI,
- age of FI,
- gender of FI,
- race/ethnicity of FI (all models except "Washington, DC, PMSA"), and
- travel status of FI.

Table 4.14 shows the effect of September 11th on SRR and IRR, both (1) unadjusted and (2) after adjusting for known correlates of nonresponse. The full results of each logistic model may be found in ***Tables 4.15*** to ***4.22***.

Screening Response Rate. In the full national sample, there was a significant decrease in SRR following September 11th. This difference remained even when controlling on all variables ($p < 0.001$; ***Table 4.15***). Likewise, in the New York City CMSA and the "Nation Excluding New York City and Washington, DC," models, the decline following September 11th remained when controlling on all factors ($p = 0.010$ and $p < 0.001$, respectively; see ***Tables 4.16*** and ***4.17***). As previously discussed, the two-sided test of differences in SRR for the Washington, DC, PMSA was marginally significant ($d = -5.34$, $p = 0.048$; see ***Table 4.3***). When controlling on all variables, the effect of September 11th was not significant at the 0.05 level ($p = 0.065$; ***Table 4.18***).

[10] Age of FI was missing for approximately 6 percent of all screening and interviewing cases. A "missing age" category was added to that variable in order to prevent the cases from being discarded from the analysis.

[11] Race/ethnicity of FI was missing for less than 0.5 percent of all screening and interviewing cases. Race/ethnicity of FI was not missing for any cases in the New York City CMSA or the Washington, DC, PMSA; it was not included in models for the Washington, DC, PMSA because all FIs in that area were of the same race/ethnicity. Because missingness was so low, a separate "missing race/ethnicity" category was not added to the variable; the missing cases were discarded.

Table 4.14 Odds Ratios Showing September 11th Effect for Weighted Screening and Interview Response Rates, 2000-2001 NSDUH

	Nation	Nation Excluding NYC CMSA and DC PMSA	NYC CMSA	DC PMSA
Screening Response Rate (SRR)				
Before Adjustment[1]	0.80[a]	0.83[a]	0.60[a]	0.56[a]
After Adjustment[2]	0.79[a]	0.83[a]	0.73[a]	0.68
Interview Response Rate (IRR)				
Before Adjustment	0.91[a]	0.93	0.73	0.63
After Adjustment	0.90[a]	0.93	0.73	0.54[a]

NYC CMSA = New York City consolidated metropolitan statistical area
DC PMSA = Washington, DC, primary metropolitan statistical area
[a] Statistically significant at the 0.05 level.
[1]Before adjustment = two-sided test of differences (see ***Tables 4.3*** and ***4.4***).
[2]After adjustment = logistic regression model (see ***Tables 4.15*** to ***4.22***).

Clearly, the effects of the September 11th tragedy had an impact on screening even after controlling for known correlates of nonresponse. This supports the hypotheses that restricted travel and other physical barriers, as well as the difficulty of making contact with and persuading respondents, would combine to make screenings much more difficult to complete.

Interview Response Rate. Controlling on the various demographic factors in the models somewhat affects the results of the IRR analyses. In the full national sample, the significant but minor difference between pre- and post-September 11th response rates ($p = 0.036$; ***Table 4.4***) remains significant ($p = 0.020$; ***Table 4.19***) when controlling on the factors listed above. The response rate models for the New York City CMSA and the "Nation Excluding New York City and Washington, DC," areas show that there was no effect of September 11th when controlling on all other factors ($p = 0.122$; ***Table 4.20*** and $p = 0.144$; ***Table 4.21***). The difference test between Q4 2000 and Q4 2001 for the Washington, DC, PMSA showed no significant difference in response rates ($d = -6.95$, $p = 0.137$; ***Table 4.4***). Interestingly, when controlling on all of the variables, this difference became significant ($p = 0.045$; ***Table 4.22***). The response rate modeling enabled detection of a change that had not been evident in the simpler analysis to be detected. This might indicate that the events of September 11th had opposite effects on different subpopulations in the Washington, DC, area. In the simpler *t* test, these effects would have cancelled each other out, but when the demographic variables were held constant, the "September 11th" effect would have been more evident.

In summary, unlike the patterns seen in SRR, the effects of September 11th on IRR appear to have been mostly indirect. The apparent tendency of some people to refuse earlier in the process (i.e., at the screening stage) might have meant that the people who did participate in the screening were more disposed to participation in general. It would seem that if there were any effects of September 11th on the IRR, they actually manifested in the screening stage. Once the interviewer was successful in contacting and screening the household, the process was apparently only slightly influenced by the aftereffects of the terrorist attacks. The only exception to this is in the Washington, DC, PMSA, which showed a significant effect of September 11th on IRR (see ***Table 4.22***).

Table 4.15 Screening Response Rate Model–Nation, Quarter 4 2000 and Quarter 4 2001 NSDUH

		Odds Ratio	95 Percent Confidence Limits	*P* Value
Intercept		19.52	(15.20, 25.06)	<0.001
September 11, 2001	Before	1.00	(1.00, 1.00)	.
	After	0.79	(0.73, 0.85)	<0.001
Concentration Non-Hispanic Black in Segment	<10%	0.99	(0.81, 1.20)	0.890
	10%-49%	0.91	(0.78, 1.05)	0.181
	≥50%	1.00	(1.00, 1.00)	.
Concentration Owner-Occupied DUs in Segment	<10%	1.47	(1.23, 1.76)	<0.001
	10%-49%	1.14	(0.92, 1.41)	0.223
	≥50%	1.00	(1.00, 1.00)	.
Concentration Hispanic in Segment	<20%	2.18	(1.38, 3.43)	0.001
	20%-70%	1.49	(1.26, 1.76)	<0.001
	≥71%	1.00	(1.00, 1.00)	.
Population Density	MSA ≥1 million	0.57	(0.51, 0.64)	<0.001
	MSA <1 million	0.71	(0.63, 0.80)	<0.001
	Non-MSA	1.00	(1.00, 1.00)	
Census Region	Northeast	0.60	(0.53, 0.68)	<0.001
	Midwest	0.82	(0.72, 0.93)	0.002
	South	0.85	(0.74, 0.97)	0.016
	West	1.00	(1.00, 1.00)	.
Number of Screenings Completed by FI	0-59	0.77	(0.06, 0.90)	0.001
	60-119	0.86	(0.74, 1.01)	0.058
	120-179	0.93	(0.81, 1.07)	0.324
	180-299	0.97	(0.88, 1.07)	0.538
	≥300	1.00	(1.00, 1.00)	.
Age of FI	0-30	0.99	(0.79, 1.24)	0.903
	31-40	1.01	(0.88, 1.16)	0.879
	41-50	1.00	(1.00, 1.00)	.
	51-60	0.95	(0.86, 1.07)	0.404
	>60	1.09	(0.96, 1.25)	0.198
	n/a	1.14	(0.96, 1.35)	0.142
Gender of FI	Male	0.95	(0.86, 1.05)	0.339
	Female	1.00	(1.00, 1.00)	.
Race/Ethnicity of FI	White, non-Hispanic	1.00	(1.00, 1.00)	.
	Black, non-Hispanic	0.89	(0.78, 1.02)	0.105
	Hispanic	0.76	(0.61, 0.94)	0.010
	Other	0.82	(0.67, 0.99)	0.041
Travel Status of FI	TFI	0.84	(0.65, 1.08)	0.165
	Non-TFI	1.00	(1.00, 1.00)	.

DU = dwelling unit; FI = field interviewer; MSA = metropolitan statistical area; TFI = traveling field interviewer; n/a = not available.

Table 4.16 Screening Response Rate Model—New York City CMSA, Quarter 4 2000 and Quarter 4 2001 NSDUH

		Odds Ratio	95 Percent Confidence Limits	*P* Value
Intercept		5.07	(3.34, 7.71)	<0.001
September 11, 2001	Before	1.00	(1.00, 1.00)	.
	After	0.73	(0.57, 0.92)	0.010
Concentration Non-Hispanic Black in Segment	<10%	1.05	(0.68, 1.61)	0.841
	10%-49%	0.89	(0.61, 1.31)	0.564
	≥50%	1.00	(1.00, 1.00)	.
Concentration Owner-Occupied DUs in Segment	<10%	1.97	(1.34, 2.90)	0.001
	10%-49%	1.06	(0.75, 1.50)	0.728
	≥50%	1.00	(1.00, 1.00)	.
Concentration Hispanic in Segment	<20%	2.84	(0.99, 8.13)	0.051
	20%-70%	1.61	(1.15, 2.25)	0.005
	≥71%	1.00	(1.00, 1.00)	.
Number of Screenings Completed by FI	0-59	0.85	(0.61, 1.19)	0.353
	60-119	1.05	(0.78, 1.41)	0.768
	120-179	0.86	(0.64, 1.16)	0.323
	180-299	0.96	(0.77, 1.19)	0.683
	≥300	1.00	(1.00, 1.00)	.
Age of FI	0-30	0.48	(0.28, 0.83)	0.009
	31-40	0.98	(0.71, 1.34)	0.891
	41-50	1.00	(1.00, 1.00)	.
	51-60	0.57	(0.44, 0.73)	<0.001
	>60	1.27	(0.82, 1.98)	0.287
	n/a	1.53	(0.87, 2.68)	0.138
Gender of FI	Male	1.14	(0.89, 1.45)	0.295
	Female	1.00	(1.00, 1.00)	.
Race/Ethnicity of FI	White, non-Hispanic	1.00	(1.00, 1.00)	.
	Black, non-Hispanic	1.10	(0.73, 1.65)	0.654
	Hispanic	0.83	(0.58, 1.19)	0.310
	Other	0.59	(0.38, 0.91)	0.016
Travel Status of FI	TFI	0.73	(0.50, 1.08)	0.140
	Non-TFI	1.00	(1.00, 1.00)	.

DU = dwelling unit; FI = field interviewer; CMSA = consolidated metropolitan statistical area; TFI = traveling field interviewer; n/a = not available.

Table 4.17 Screening Response Rate Model—Nation Excluding the New York City CMSA and the Washington, DC, PMSA, Quarter 4 2000 and Quarter 4 2001 NSDUH

		Odds Ratio	95 Percent Confidence Limits	*P* Value
Intercept		19.07	(14.62, 24.87)	<0.001
September 11, 2001	Before	1.00	(1.00, 1.00)	.
	After	0.83	(0.76, 0.91)	<0.001
Concentration Non-Hispanic Black in Segment	<10%	0.99	(0.80, 1.24)	0.949
	10%-49%	0.92	(0.79, 1.08)	0.311
	≥50%	1.00	(1.00, 1.00)	.
Concentration Owner-Occupied DUs in Segment	<10%	1.27	(1.04, 1.56)	0.019
	10%-49%	1.08	(0.85, 1.36)	0.541
	≥50%	1.00	(1.00, 1.00)	.
Concentration Hispanic in Segment	<20%	2.04	(1.23, 3.39)	0.006
	20%-70%	1.56	(1.28, 1.91)	<0.001
	≥71%	1.00	(1.00, 1.00)	.
Population Density	MSA ≥1 million	0.60	(0.53, 0.67)	<0.001
	MSA <1 million	0.70	(0.62, 0.79)	<0.001
	Non-MSA	1.00	(1.00, 1.00)	.
Census Region	Northeast	0.75	(0.66, 0.86)	<0.001
	Midwest	0.86	(0.76, 0.98)	0.020
	South	0.90	(0.79, 1.03)	0.122
	West	1.00	(1.00, 1.00)	.
Number of Screenings Completed by FI	0-59	0.78	(0.66, 0.92)	0.003
	60-119	0.88	(0.74, 1.04)	0.138
	120-179	0.97	(0.84, 1.13)	0.729
	180-299	0.99	(0.88, 1.10)	0.790
	≥300	1.00	(1.00, 1.00)	.
Age of FI	0-30	1.06	(0.84, 1.33)	0.629
	31-40	1.01	(0.87, 1.17)	0.908
	41-50	1.00	(1.00, 1.00)	.
	51-60	1.02	(0.91, 1.14)	0.789
	>60	1.10	(0.96, 1.27)	0.181
	n/a	1.10	(0.91, 1.32)	0.344
Gender of FI	Male	0.93	(0.83, 1.04)	0.208
	Female	1.00	(1.00, 1.00)	.
Race/Ethnicity of FI	White, non-Hispanic	1.00	(1.00, 1.00)	.
	Black, non-Hispanic	0.87	(0.75, 1.02)	0.078
	Hispanic	0.87	(0.68, 1.10)	0.244
	Other	0.91	(0.74, 1.12)	0.369
Travel Status of FI	TFI	0.98	(0.70, 1.37)	0.914
	Non-TFI	1.00	(1.00, 1.00)	.

DU = dwelling unit; FI = field interviewer; MSA = metropolitan statistical area; CMSA = consolidated metropolitan statistical area; PMSA = primary metropolitan statistical area; TFI = traveling field interviewer; n/a = not available.

Table 4.18 Screening Response Rate Model—Washington, DC, PMSA, Quarter 4 2000 and Quarter 4 2001 NSDUH

		Odds Ratio	95 Percent Confidence Limits	*P* Value
Intercept		7.43	(3.86, 14.33)	<0.001
September 11, 2001	Before	0.68	(1.00, 1.00)	.
	After	0.68	(0.45, 1.03)	0.065
Concentration Non-Hispanic Black in Segment	<10%	0.97	(0.50, 1.91)	0.939
	10%-49%	0.59	(0.28, 1.22)	0.152
	≥50%	1.00	(1.00, 1.00)	.
Concentration Owner-Occupied DUs in Segment	<10%	1.31	(0.61, 2.79)	0.491
	10%-49%	0.88	(0.40, 1.94)	0.755
	≥50%	1.00	(1.00, 1.00)	.
Concentration Hispanic in Segment	<20%	1.85	(0.78, 4.39)	0.163
	20%-70%	1.00	(1.00, 1.00)	.
Number of Screenings Completed by FI	0-59	1.03	(0.55, 1.92)	0.919
	60-119	0.59	(0.26, 1.31)	0.195
	120-179	0.83	(0.38, 1.80)	0.633
	180-299	2.72	(1.35, 5.50)	0.005
	≥300	1.00	(1.00, 1.00)	
Age of FI	0-30	1.26	(0.36, 4.40)	0.715
	31-40	1.14	(0.41, 3.15)	0.800
	41-50	1.00	(1.00, 1.00)	.
	51-60	2.00	(0.95, 4.17)	0.066
	>60	1.58	(0.48, 5.20)	0.450
	n/a	1.05	(0.42, 2.66)	0.914
Gender of FI	Male	0.94	(0.37, 2.35)	0.890
	Female	1.00	(1.00, 1.00)	.
Travel Status of FI	TFI	0.36	(0.21, 0.03)	<0.001
	Non-TFI	1.00	(1.00, 1.00)	.

DU = dwelling unit; FI = field interviewer; PMSA = primary metropolitan statistical area; TFI = traveling field interviewer; n/a = not available.

Table 4.19 Interview Response Rate Model—Nation, Quarter 4 2000 and Quarter 4 2001 NSDUH

		Odds Ratio	95 Percent Confidence Limits	*P* Value
Intercept		2.30	(1.73, 3.05)	<0.001
September 11, 2001	Before	1.00	(1.00, 1.00)	.
	After	0.90	(0.83, 0.98)	0.020
Race/Ethnicity of Respondent	White, non-Hispanic	1.45	(1.16, 1.82)	0.001
	Black, non-Hispanic	1.51	(1.17, 1.96)	0.002
	Hispanic	1.84	(1.43, 2.39)	<0.001
	Other	1.00	(1.00, 1.00)	.
Age of Respondent	12-17	2.08	(1.85, 2.33)	<0.001
	18-25	1.36	(1.23, 1.52)	<0.001
	26-34	1.29	(1.13, 1.46)	<0.001
	35-49	1.22	(1.08, 1.37)	0.001
	≥50	1.00	(1.00, 1.00)	.
Gender of Respondent	Male	0.90	(0.82, 0.98)	0.017
	Female	1.00	(1.00, 1.00)	.
Population Density	MSA ≥1 million	0.75	(0.67, 0.85)	<0.001
	MSA <1 million	0.83	(0.74, 0.94)	0.003
	Non-MSA	1.00	(1.00, 1.00)	.
Census Region	Northeast	0.88	(0.76, 1.02)	0.008
	Midwest	0.97	(0.85, 1.11)	0.702
	South	1.04	(0.90, 1.21)	0.561
	West	1.00	(1.00, 1.00)	.
Number of Screenings Completed by FI	0-59	0.75	(0.64, 0.87)	<0.001
	60-119	0.88	(0.76, 1.03)	0.109
	120-179	0.89	(0.76, 1.04)	0.150
	180-299	0.85	(0.77, 0.95)	0.006
	≥300	1.00	(1.00, 1.00)	.
Age of FI	0-30	1.20	(0.95, 1.53)	0.124
	31-40	0.93	(0.82, 1.06)	0.288
	41-50	1.00	(1.00, 1.00)	.
	51-60	0.93	(0.82, 1.05)	0.244
	>60	0.97	(0.84, 1.11)	0.649
	n/a	1.05	(0.85, 1.30)	0.635
Gender of FI	Male	1.03	(0.92, 1.16)	0.568
	Female	1.00	(1.00, 1.00)	.
Race/Ethnicity of FI	White, non-Hispanic	1.00	(1.00, 1.00)	.
	Black, non-Hispanic	1.18	(0.99, 1.40)	0.060
	Hispanic	0.92	(0.75, 1.11)	0.376
	Other	1.23	(0.97, 1.57)	0.093
Travel Status of FI	TFI	1.10	(0.84, 1.43)	0.494
	Non-TFI	1.00	(1.00, 1.00)	.

FI = field interviewer; MSA = metropolitan statistical area; TFI = traveling field interviewer; n/a = not available.

Table 4.20 Interview Response Rate Model—New York City CMSA, Quarter 4 2000 and Quarter 4 2001 NSDUH

		Odds Ratio	95 Percent Confidence Limits	*P* Value
Intercept		1.31	(0.68, 2.54)	0.422
September 11, 2001	Before	1.00	(1.00, 1.00)	.
	After	0.73	(0.49, 1.09)	0.122
Race/Ethnicity of Respondent	White, non-Hispanic	1.62	(1.12, 2.34)	0.011
	Black, non-Hispanic	2.11	(1.19, 3.74)	0.011
	Hispanic	2.05	(1.21, 3.48)	0.008
	Other	1.00	(1.00, 1.00)	.
Age of Respondent	12-17	2.36	(1.45, 3.85)	0.001
	18-25	1.24	(0.85, 1.80)	0.261
	26-34	1.64	(1.06, 2.54)	0.027
	35-49	1.28	(0.82, 1.98)	0.278
	≥50	1.00	(1.00, 1.00)	.
Gender of Respondent	Male	0.98	(0.67, 1.42)	0.906
	Female	1.00	(1.00, 1.00)	.
Number of Screenings Completed by FI	0-59	1.06	(0.50, 2.26)	0.871
	60-119	1.19	(0.63, 2.24)	0.600
	120-179	1.10	(0.47, 2.58)	0.819
	180-299	0.92	(0.60, 1.46)	0.687
	≥300	1.00	(1.00, 1.00)	.
Age of FI	0-30	0.26	(0.10, 0.67)	0.005
	31-40	0.70	(0.43, 1.13)	0.147
	41-50	1.00	(1.00, 1.00)	.
	51-60	0.98	(0.63, 1.52)	0.928
	>60	1.11	(0.59, 2.08)	0.743
	n/a	1.61	(0.86, 3.01)	0.139
Gender of FI	Male	1.05	(0.69, 1.58)	0.829
	Female	1.00	(1.00, 1.00)	.
Race/Ethnicity of FI	White, non-Hispanic	1.00	(1.00, 1.00)	.
	Black, non-Hispanic	1.16	(0.73, 1.84)	0.531
	Hispanic	0.78	(0.36, 1.65)	0.508
	Other	0.80	(0.46, 1.40)	0.433
Travel Status of FI	TFI	0.99	(0.39, 2.52)	0.983
	Non-TFI	1.00	(1.00, 1.00)	.

FI = field interviewer; CMSA = consolidated metropolitan statistical area; TFI = traveling field interviewer; n/a = not available.

Table 4.21 Interview Response Rate Model—Nation Excluding New York City CMSA and Washington, DC, PMSA, Quarter 4 2000 and Quarter 4 2001 NSDUH

		Odds Ratio	95 Percent Confidence Limits	*P* Value
Intercept		2.41	(1.79, 3.26)	<0.001
September 11, 2001	Before	1.00	(1.00, 1.00)	.
	After	0.93	(0.85, 1.02)	0.144
Race/Ethnicity of Respondent	White, non-Hispanic	1.41	(1.11, 1.80)	0.005
	Black, non-Hispanic	1.47	(1.11, 1.93)	0.007
	Hispanic	1.84	(1.40, 2.43)	<0.001
	Other	1.00	(1.00, 1.00)	.
Age of Respondent	12-17	2.06	(1.83, 2.33)	<0.001
	18-25	1.37	(1.22, 1.53)	<0.001
	26-34	1.27	(1.11, 1.45)	<0.001
	35-49	1.20	(1.06, 1.35)	0.004
	≥50	1.00	(1.00, 1.00)	.
Gender of Respondent	Male	0.89	(0.81, 0.98)	0.014
	Female	1.00	(1.00, 1.00)	.
Population Density	MSA ≥1 million	0.73	(0.64, 0.83)	<0.001
	MSA <1 million	0.84	(0.74, 0.95)	0.004
	Non-MSA	1.00	(1.00, 1.00)	.
Census Region	Northeast	0.90	(0.78, 1.04)	0.144
	Midwest	0.98	(0.86, 1.12)	0.769
	South	1.01	(0.87, 1.18)	0.859
	West	1.00	(1.00, 1.00)	.
Number of Screenings Completed by FI	0-59	0.72	(0.61, 0.83)	<0.001
	60-119	0.88	(0.75, 1.03)	0.105
	120-179	0.88	(0.75, 1.02)	0.093
	180-299	0.85	(0.75, 0.95)	0.004
	≥300	1.00	(1.00, 1.00)	.
Age of FI	0-30	1.29	(1.03, 1.62)	0.029
	31-40	0.92	(0.80, 1.05)	0.202
	41-50	1.00	(1.00, 1.00)	.
	51-60	0.92	(0.81, 1.04)	0.182
	>60	0.95	(0.82, 1.10)	0.503
	n/a	1.03	(0.83, 1.29)	0.760
Gender of FI	Male	1.04	(0.92, 1.17)	0.557
	Female	1.00	(1.00, 1.00)	.
Race/Ethnicity of FI	White, non-Hispanic	1.00	(1.00, 1.00)	.
	Black, non-Hispanic	1.15	(0.96, 1.38)	0.121
	Hispanic	0.93	(0.76, 1.14)	0.493
	Other	1.25	(0.96, 1.64)	0.100
Travel Status of FI	TFI	1.06	(0.80, 1.39)	0.692
	Non-TFI	1.00	(1.00, 1.00)	.

FI = field interviewer; CMSA = consolidated metropolitan statistical area; PMSA = primary metropolitan statistical area; TFI = traveling field interviewer; n/a = not available.

Table 4.22 Interview Response Rate Model—Washington, DC, PMSA, Quarter 4 2000 and Quarter 4 2001 NSDUH

		Odds Ratio	95 Percent Confidence Limits	*P* Value
Intercept		1.30	(0.32, 5.29)	0.714
September 11, 2001	Before	1.00	(1.00, 1.00)	.
	After	0.54	(0.29, 0.99)	0.045
Race/Ethnicity of Respondent	White, non-Hispanic	2.93	(0.33, 26.03)	0.335
	Black, non-Hispanic	2.09	(0.34, 12.79)	0.423
	Hispanic	7.55	(0.87, 65.42)	0.067
	Other	1.00	(1.00, 1.00)	.
Age of Respondent	12-17	1.70	(0.69, 4.19)	0.245
	18-25	1.74	(0.79, 3.83)	0.169
	26-34	0.92	(0.39, 2.15)	0.849
	35-49	2.36	(0.71, 7.89)	0.163
	≥50	1.00	(1.00, 1.00)	.
Gender of Respondent	Male	0.78	(0.34, 1.79)	0.552
	Female	1.00	(1.00, 1.00)	.
Number of Screenings Completed by FI	0-59	1.90	(0.96, 3.77)	0.065
	60-119	2.94	(0.80, 10.80)	0.105
	120-179	3.37	(0.92, 12.40)	0.068
	180-299	1.53	(0.62, 3.78)	0.359
	≥300	1.00	(1.00, 1.00)	.
Age of FI	0-30	2.21	(0.54, 9.10)	0.271
	31-40	1.14	(0.38, 3.41)	0.816
	41-50	1.00	(1.00, 1.00)	.
	51-60	0.99	(0.41, 2.43)	0.988
	>60	2.20	(0.47, 10.29)	0.315
	n/a	1.21	(0.31, 4.77)	0.784
Gender of FI	Male	0.98	(0.49, 1.97)	0.951
	Female	1.00	(1.00, 1.00)	.
Travel Status of FI	TFI	0.55	(0.17, 1.78)	0.319
	Non-TFI	1.00	(1.00, 1.00)	.

FI = field interviewer; PMSA = primary metropolitan statistical area; TFI = traveling field interviewer; n/a = not available

Conclusions

As expected, it was found that the New York City CMSA and Washington, DC, PMSA response rates suffered dramatic decreases following the September 11th terrorist attacks, though the differences in IRR in the Washington, DC, PMSA were shown to be significant only after modeling on a number of factors. The national SRR also showed a decrease even after removing the New York City CMSA and Washington, DC, PMSA from the sample. This decrease was significant but less dramatic than in the two metropolitan areas.

One must bear in mind that the New York City sample supplement in Q4 2001 produced a potentially confounding effect. The changes in response rates in the New York City CMSA might have been a result of the increased number of cases to be worked and not a direct result of September 11th. Even though TFIs were brought in to assist with the additional work, each FI still had a significantly larger caseload than usual. This could potentially have caused less

attention to be paid to each case, which would have resulted in less success in contacting households and gaining cooperation. However, the likelihood of this effect is minimal in that some decline had already been demonstrated in Q3 and similar results were observed in the Washington, DC, PMSA. Rather, the changes in the New York City CMSA response rates appear to have been related to September 11th, and the Q4 supplement increased the sample size sufficiently for the significance of those changes to be detected. Similarly, if the sample size in the Washington, DC, PMSA had been greater, the large response rate differences seen there might have more readily been proven statistically significant.

References

CNN.com. (n.d.). *Anthrax investigation*. Retrieved June 24, 2004, from http://www.cnn.com/interactive/health/0110/anthrax/frameset.exclude.html

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: Wiley.

National Institute on Drug Abuse. (2002, March 27). *Study assesses impact of September 11th events on Manhattan residents: Participants report symptoms of post traumatic stress disorder and depression* [news release]. Retrieved June 24, 2004, from http://www.drugabuse.gov/MedAdv/02/NR3-27.html

Office of Applied Studies. (2002). *Impact of September 11, 2001 events on substance use and mental health* (DHHS Publication No. SMA 02-3729, Analytic Series A-18). Rockville, MD: Substance Abuse and Mental Health Services Administration. [Available at http://www.oas.samhsa.gov/analytic.htm]

Research Triangle Institute. (2001). *SUDAAN user's manual: Release 8.0*. Research Triangle Park, NC: Author.

Schlenger, W. E., Caddell, J. M., Ebert, L., Jordan, B. K., Rourke, K. M., Wilson, D., Thalji, L., Dennis, J. M., Fairbank, J. A., & Kulka, R. A. (2002). Psychological reactions to terrorist attacks: Findings from the National Study of Americans' Reactions to September 11. *Journal of the American Medical Association, 288*, 581-588.

Schuster, M. A., Stein, B. D., Jaycox, L., Collins, R. L., Marshall, G. N., Elliott, M. N., Zhou, A. J., Kanouse, D. E., Morrison, J. L., & Berry, S. H. (2001). A national survey of stress reactions after the September 11, 2001, terrorist attacks. *New England Journal of Medicine, 345*, 1507-1512.

Sheatsley, P. B., & Feldman, J. J. (1964). The assassination of President Kennedy: A preliminary report on public reactions and behavior. *Public Opinion Quarterly, 28*, 189-215.

Silver, R. C., Holman, E. A., McIntosh, D. N., Poulin, M., & Gil-Rivas, V. (2002). Nationwide longitudinal study of psychological responses to September 11. *Journal of the American Medical Association, 288*, 1235-1244.

Smith, D. W., Christiansen, E. H., Vincent, R., & Hann, N. E. (1999). Population effects of the bombing of Oklahoma City. *Journal – Oklahoma State Medical Association, 92*(4), 193-198.

Smith, T. W., & Jarkko, L. (1998, May). *National pride: A cross-national analysis* (GSS Cross National Report No. 19). Chicago, IL: National Opinion Research Center. [Available as a PDF at http://www.norc.uchicago.edu/new/part1.pdf]

Smith, T. W., Rasinski, K. A., & Toce, M. (2001, October 25). *America rebounds: A national study of public response to the September 11th terrorist attacks (preliminary findings)*. Chicago, IL: National Opinion Research Center. [Available as a PDF at http://www.norc.uchicago.edu/projects/reaction/pubresp.pdf]

5. Association between Interviewer Experience and Substance Use Prevalence Rates in NSDUH

James R. Chromy, Joe Eyerman, Dawn Odom,[1] and Madeline E. McNeeley
RTI International

Art Hughes
Substance Abuse and Mental Health Services Administration

Introduction

Analysis of survey data from the National Survey on Drug Use and Health (NSDUH) has shown a relationship between interviewer experience, response rates, and the prevalence of self-reported substance use (Eyerman, Odom, Wu, & Butler, 2002; Hughes, Chromy, Giacoletti, & Odom, 2001, 2002; Office of Applied Studies [OAS], 2002a, 2002b). These analyses have shown a significant and positive relationship between the amount of prior experience an interviewer has with collecting NSDUH data and the response rates that the interviewer produces with his or her workload. The analyses also have shown a significant and negative relationship between the amount of prior experience of an interviewer and the prevalence of substance use reported in cases completed by that interviewer. This chapter describes the methodology employed to explain these effects within a unified theoretical framework.

The prior analyses mentioned above examined interviewer response rates and prevalence rates independently. This has made it difficult to determine if the lower prevalence rates for experienced interviewers are a result of the change in the sample composition due to higher response rates or if the lower prevalence rates are a result of a direct effect of interviewer behavior on respondent self-reporting. This study combines these two explanations to produce a conceptual model that summarizes the expectations for the relationship between interviewer experience and prevalence rates. The combined explanation from the conceptual model is evaluated in a series of conditional models to examine the indirect effect of response rates and the direct effect of interviewer experience on prevalence rates.

The NSDUH design is a multistage area probability sample that targets a respondent universe of noninstitutionalized civilians aged 12 or older within the 50 States and the District of Columbia. Although the survey is conducted annually, the household sample is selected and fielded quarterly. Household screening and interview respondent selection procedures are conducted with a handheld computer. If a respondent is selected and agrees to participate, the NSDUH questionnaire then is administered using computer-assisted interviewing (CAI), with an audio computer-assisted self-interviewing component (ACASI) for the more sensitive questions.

[1] Now at Inveresk.

The remainder of the interview is administered using computer-assisted personal interviewing (CAPI). These procedures are explained in order to give the reader an idea as to the points during the course of the survey at which an interviewer is most likely to have an influence on cooperation and responses.

Conceptual Model

A considerable amount of research has been conducted on the relationship between the behavior of field interviewers (FIs) and the data collected through surveys. Some of this literature has focused on the relationship between FI characteristics and the population estimates generated by the survey (Cleary, Mechanic, & Weiss, 1981; Hughes et al., 2001; Stevens & Bailar, 1976). In general, this literature demonstrates that FIs can influence both the success of the data collection process and the accuracy of population estimates.

Consistent with the literature, a series of analyses of NSDUH data have shown a strong relationship between FI characteristics and the data yielded by the survey process, including the estimates of drug use prevalence in the U.S. population. These findings have been supported with anecdotal evidence gathered through conversations with veteran project and field management staff. As expected by the literature and conventional wisdom, the anecdotal evidence suggests that the experienced interviewers have tailored their behaviors to be more productive, efficient, and persuasive.

Many of these behaviors also could affect prevalence estimates. For example, NSDUH field management has noted that interviewers tend to be reluctant to use the reference calendar, a tool intended to assist respondents with their episodic recall. The more experienced interviewers recognize that completing the calendar in the method required by the survey design can delay the completion of the interview. Therefore, they often complete the calendar incorrectly or not at all in order to expedite the interview. This may lead to lower data collection costs and higher response rates, which will lead to better performance reviews for the interviewer. However, deviations from the design by the interviewer could lead to worse recall and possibly lower prevalence estimates.

Why a Conceptual Model?

The conceptual model is a tool to organize the current understanding of the relationship around a logical structure. It allows the number of possible explanations to be reduced by removing the redundant or logically impossible. It guides the generation of research questions by identifying conflicting but equally reasonable explanations. It also generates hypotheses that can be readily tested by identifying the critical interviewer behaviors, the stage of the survey process at which the behaviors affect the prevalence rates, the best measures of those behaviors, and the expected impact of those behaviors on the prevalence rates.

NSDUH Conceptual Model

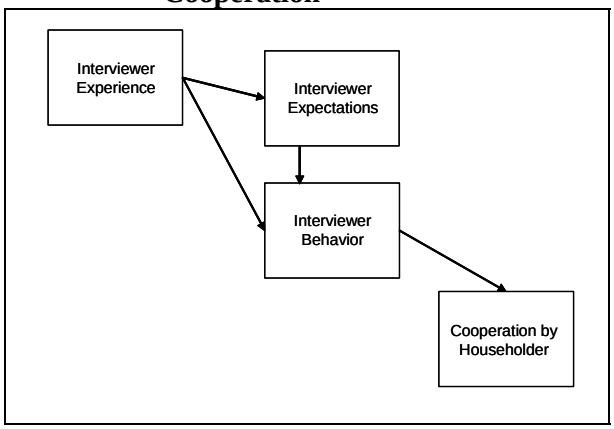
It important to recognize that the FI can affect the data collection process at both the screening and the interview stages of the survey. For example, the interviewer may manipulate the eligibility rule during the screening stage to systematically remove dwelling units with a trait

that he or she believes indicates they will be hard to complete. This will improve his or her response rates and costs per interview but also will affect prevalence rates if the trait is correlated with self-reported substance use. The same interviewer can affect the prevalence at the interview stage. For example, if he or she follows all the project protocols and is skilled at contacting and gaining cooperation, he or she will reduce nonresponse error and generate more representative estimates. However, he or she also may manipulate the protocol to gain cooperation or reduce the time required to conduct the survey, which may lead to greater measurement error.

For this reason, the conceptual model was organized around the structure of the NSDUH interview. In addition, this was done in order to be able easily to draw the existing literature on interviewer behaviors into the model. Although some of this literature addresses population estimates, most of it is organized around response rates and gaining cooperation. Therefore, it was decided to expand on the Groves and Couper (1998) conceptual model of the relationship between interviewer behaviors and cooperation.

Groves and Couper (1998) noted that interviewer experience has long been considered a critical factor in gaining the participation of sample members. ***Figure 5.1*** depicts the interviewer experience pathway of the larger model that Groves and Couper proposed. Groves and Couper posited that increased interviewer experience leads to increases in various kinds of knowledge gained on the job, which in turn lead to gains in respondent cooperation (see ***Table 5.1***). Their work is the most comprehensive attempt by the survey community thus far to develop a comprehensive view of the effects of interviewer experience.

Figure 5.1 Interviewer Experience Pathway of the Model "Interviewer Influences on Survey Cooperation"



Source: Groves and Couper (1998).

Table 5.1 Interviewer Behaviors Affected by Experience Level and Their Impact on Sample Member Cooperation

Effect of Increased Experience on Behavior	Description	Effect of Behavior on Cooperation
Credible agents	Knowledge of subject area; understanding the concerns and subject-specific "language" of the respondents	Increases cooperation rates
Dealing with resistance in the field	Knowledge of the interviewing task and confidence in dealing with respondents; ethnic/racial and local/regional knowledge	Increases cooperation rates
Tailoring approaches for gaining cooperation	Optimizing strategies based on exposure to a variety of circumstances, including techniques/guidelines of multiple surveys and organizations	Increases cooperation rates

Source: Groves and Couper (1998).

Figures 5.2 and ***5.3*** expand the Groves and Couper model to include the impact of interviewer behavior on prevalence rates. It should be noted that the *interviewer expectations* stage was removed from the model for three reasons. First, all of the influence of the *interviewer expectations* passes through the *interviewer behavior* stage of the model. Therefore, any relevant change in expectations should be observed by changes in behaviors. Second, measures of interviewer expectations are unavailable, so there is no need for a proxy concept for interviewer behavior. Third, all of the current explanations about the interviewer effect on prevalence rates are based on interviewer behavior.

Figure 5.2 Interviewer Influences on Prevalence Rates During Screening

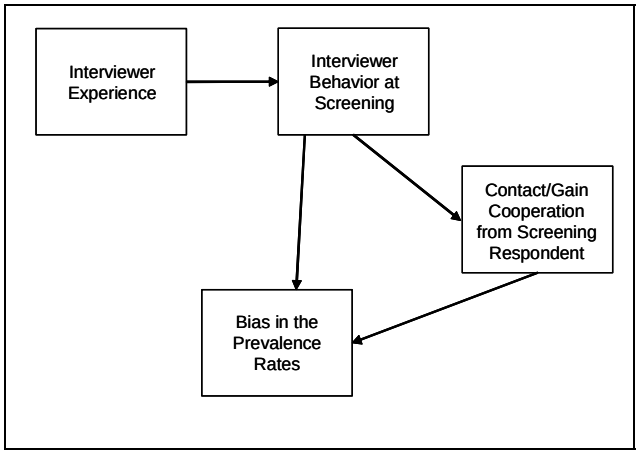


Figure 5.3 Interviewer Influences on Prevalence Rates During Interviewing

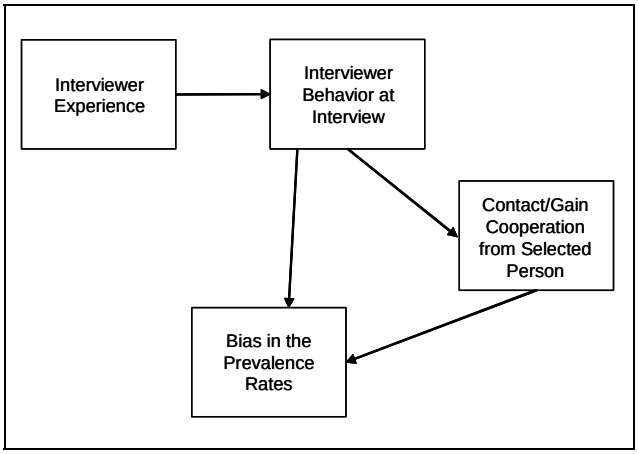
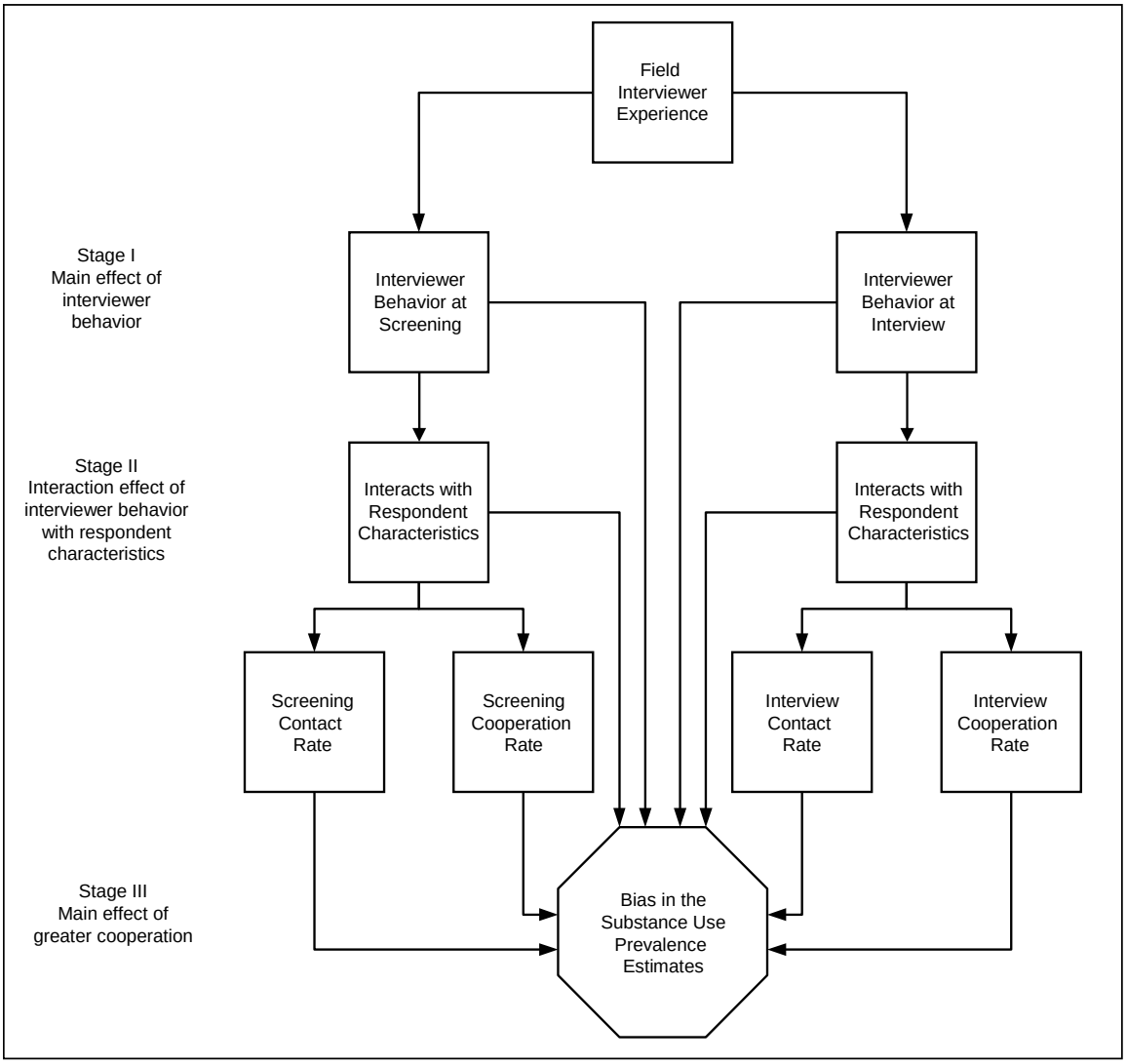


Figure 5.4 combines ***Figures 5.2*** and ***5.3*** into a single model to show the relationship between interviewer experience and the prevalence rates for the full process of screening and interviewing.

The specific behaviors to be considered by this model are explained in more detail in ***Tables 5.2*** and ***5.3***. The first set of behaviors is indirect. That is, the increase or decrease of bias introduced in the prevalence rates comes through some intermediate step (e.g., through increasing cooperation rates) (see ***Table 5.2***).

The second set of behaviors shows the direct effects of experience on prevalence estimates. As noted above, estimates can be positively influenced by the interviewer's ability to serve as a "credible agent" for the survey request. His or her knowledge of the interview and ability to communicate effectively with the respondent can significantly decrease the effects of social desirability, which will reduce the bias on the prevalence estimate. At the same time, however, interviewers might use their knowledge to manipulate the respondents' awareness of the protocols in order to gain cooperation. The lack of a full awareness of protocols could increase respondents' social desirability concerns and increase the bias on the estimate. Other effects of experience can be seen in every aspect of the interview, from establishing a good rapport with respondents to misusing or omitting the pillcards from the interview. ***Table 5.3*** summarizes interviewer behaviors that could be influenced by interviewer experience and suggests the direct effects they might have on prevalence estimates.

Figure 5.4 Model of the Paths of Influence of Field Interviewer Experience on Prevalence Estimates in the National Survey on Drug Use and Health



In ***Table 5.3***, specified changes in bias are accompanied by a suggested direction for the change in the prevalence estimate. This is because of a generally accepted tenet of survey research: When inquiring into sensitive, socially undesirable subjects, more reporting is better. Biemer (1988, 2001) described the potential fallacy in this assumption. Specifically, he pointed out that tendencies to under- or overreport might not be consistent across all population groups, particularly because the social desirability of the behavior might vary; also, a number of biases related to recall issues, nonresponse, and other factors all can interfere with the estimate. One must be cautious in the approach to this stage of the model and consider carefully any hypotheses regarding changes in the levels of the prevalence estimate.

Table 5.2 Interviewer Behaviors Affected by Experience Level and Their Indirect Impact on Prevalence Rates

Behavior	Description	Effect of Increased Experience on Cooperation Rates	Effect on Bias in Prevalence Estimates
Serving as a "credible agent" for the survey request	Knowledge of subject area; understanding the concerns and subject-specific "language" of the respondents[a]	Increase cooperation rates by increasing numerator (obtain more completes)	Decrease
Dealing with resistance in the field and obtaining cooperation from respondents	Knowledge of the interviewing task; confidence in dealing with respondents and gatekeepers; ethnic/racial and local/regional knowledge[a]	Increase cooperation rates by increasing numerator	Decrease
Tailoring approaches for making contact and gaining cooperation	Optimizing strategies based on exposure to a variety of circumstances, including techniques/guidelines of multiple surveys and organizations[a]	Increase cooperation rates by increasing numerator	Decrease
Reducing workload by reducing eligibility	Manipulating college student eligibility rule by visiting residence halls at times that will minimize eligibility; coding by observation/assumption, which results in deliberate and/or unintentional mistakes that reduce eligibility	Should increase cooperation rates by decreasing denominator (fewer sample members considered eligible to participate)	Increase
Using telephone for prescreening of SDUs	Call SDUs to screen out vacancies, businesses, etc; warn reluctant respondents of the upcoming visit	Should decrease cooperation rates by decreasing numerator (more refusals)	Increase

SDU = sample dwelling unit.
[a] Groves and Couper (1998).

Statistical Analysis

Using the conceptual model as a guideline, statistical models were created to evaluate the relationship between FI-level measures and response/prevalence rates, while controlling for demographic and other available covariates. FI experience was investigated in a series of separate models that were conditionally based on the sequence of screening and interviewing events that is portrayed in ***Figure 5.4***. This allowed for the exploration of the interrelationships of FI experience at different levels of the interviewing process. To tie each step together, the weight in each model was adjusted using response propensities based on the previous models. Although each step of the interviewing process is of interest, the ultimate goal is to see how prevalence rates are affected by interviewer experience given each of the previous interviewing steps. ***Table 5.4*** contains a detailed description of the creation of the datasets and the statistical models.

Description of Variables and Datasets

In an attempt to explain the effect of FI experience on response and prevalence rates, the available data were collected and transformed into usable datasets that could be analyzed. The screening and interview data can be grouped into the conceptual framework based on Groves and Couper's model (1998). As described later, the framework explains the process of survey participation as the interaction between the respondent and the interviewer, which is influenced by four factors: social environment, householder (respondent) characteristics, interviewer characteristics, and survey design. When investigating how respondents report drug use, a model

Table 5.3 Interviewer Behaviors Affected by Experience Level and Their Direct Impact on Prevalence Rates

Source	Behavior	Description	Effect of Increased Experience on Interview Process	Effect on Bias in Prevalence Estimates
Behavior from Groves and Couper (1998); effects described are authors' own	Serving as a "credible agent" for the survey request and dealing with resistance	Knowledge of subject area; understanding the concerns and subject-specific "language" of the respondents; knowledge of the interviewing task	Effective communication and professionalism; decreases social desirability concerns	Decrease
			Manipulates cooperation by decreasing respondent awareness; increases social desirability concerns	Increase
Körmendi and Pedersen (1995); Cleary et al. (1981)	Atmosphere/rapport	Comfortable with asking sensitive questions and able to create a relaxed, friendly atmosphere	Decreases social desirability concerns; encourages thorough and thoughtful answers	Decrease
	Failing to ensure privacy of interview	Value a completed interview over ensuring the privacy of that interview, especially for child interviews	Increases social desirability concerns	Increase
	Neglecting the keyboard tutorial	Desire to hurry the interview causes interviewers to rush the tutorial and fail to explain the function keys thoroughly—important because the DK/REF keys are not explained in the ACASI	Lack of awareness of DK/REF keys will decrease item nonresponse; could result in more correct answers as well as more false answers	Increase
Stevens and Bailar (1976)	Completes all parts of interview more quickly—"rate"—and body language	Makes respondent feel pressured to move quickly	Recall is worse (more likely to telescope events out of the reference period)	Increase
	Coaching	Tells survey respondent that answering "no" to gate questions makes interview go more quickly	Increases number of false "no" answers	Increase
	Calendar	Omits calendar, fills out carelessly, or gives little instruction/encouragement	Recall is worse (may telescope events into or out of the reference period, depending on type of calendar misuse)	Increase
	Headphones	Experienced are more likely to use because it is second nature	Improves understanding of questions	—
	Pillcards	Experienced do not use, use casually, hand book over to respondent to use on his or her own	Recall is worse	Increase

ACASI = audio computer-assisted self-interviewing; DK/REF = don't know/refused.

Table 5.4 Series of Conditional Models Investigating FI Experience

Model	Dependent Variable	Data file	Base Weight[a]
1	Contacting HH	Selected households	Household (design-based)[b]
2	Gaining HH cooperation (i.e., successful screening)	Selected households	Household (design-based)[b]
3	Contacting selected person in HH	Selected persons	Person (design-based)[c]
4	Successfully interviewing selected person	Selected persons	Person (design-based)[c]
5	Responding person reports substance use	Completed person	Person (design-based)[c]

HH = household.

[a] The weights in each step (excluding the first step) also include propensity adjustments derived from the previous model steps.

[b] The household (design-based) weight does not include a household nonresponse, poststratification or extreme weight adjustment.

[c] The person (design-based) weight includes all household weights but does not include a person nonresponse, poststratification, or extreme weight adjustment.

of substance use was used based on past NSDUH experience.[2] The data sources used for the analysis include segment-level census data, FI demographics, design characteristics, selected person characteristics collected during the screening stage of the interview, and respondent interview data.

Unfortunately, other measures also were available but were excluded due to limitations in scope and coverage. For example, observations of the FIs provided a rich source of data but were administered only to a subset of the interviewers and only in 2001. Verification data also provide insight into how the interviewer actually administered the interview; however, these data also were collected only on a small subset of interviewers. Analysis of these data provides a considerable detail in specific situations but cannot be used to generalize to the full project because they apply only to a subset. For this reason, the set of analyses reviewed in this document is limited to those measures that can be applied to the entire project for 1999, 2000, and 2001.

In general, the variables included in the model were consistent with Groves and Couper (1998) for predictors of survey participation and previous NSDUH experience for predictors of substance use.

- **Environment**: Census region, population density, segment-level characteristics (based on updated census data) including percentage Hispanic concentration, percentage non-Hispanic black population, and percentage of households that are owner-occupied.
- **Respondent Characteristics**: Selected person race/ethnicity, gender, and age category (where available).
- **Interviewer Characteristics**: Race/ethnicity, gender, and age category of interviewer. Interviewer experience is defined in two ways. The first way reports the number of *screenings* an interviewer has conducted since 1999 and is classified into the following three categories:

Inexperienced	=	0 to 119 screenings since 1999 (for 1999 survey: no experience prior to 1999)
Experienced	=	120 to 299 screenings since 1999 (for 1999 survey: and/or experience prior to 1999)
Highly Experienced	=	300+ screenings since 1999

  The second way reports the number of *interviews* an interviewer has conducted since 1999 and is classified into following three categories:

Inexperienced	=	0 to 39 interviews since 1999 (for 1999 survey: no experience prior to 1999)
Experienced	=	40 to 99 interviews since 1999 (for 1999 survey: and/or experience prior to 1999)
Highly Experienced	=	100+ interviews since 1999

[2] Prior to 2002, the survey was called the National Household Survey on Drug Abuse (NHSDA). In this report, it is referred to as NSDUH, regardless of the survey year.

- **Survey Design**: Survey year, number of persons selected in household, and whether segment was included in the 2001 incentive experiment (IE),[3] which is classified into the following three categories:
  1 = Not an IE FI region
  2 = IE FI region but did not receive incentive payment
  3 = IE FI region and did receive incentive payment

The following datasets were created for the analysis and include, where available, the variables described above:

- **Selected Household-Level File**: A household-level file was created for Models 1 and 2 that includes every selected household from 3 survey years: 1999, 2000, and 2001. Because information is limited at the time that screening occurs, only variables focusing on the characteristics of the FI and environmental characteristics (i.e., primary sampling unit/segment characteristics) were available. Because several FIs could possibly conduct screening and interviewing (S&I) on the same household, it was decided to associate the first FI who visited the household as the FI who conducted the entire S&I. (Note: This does not penalize refusal-converters, and at least 75 percent of the households have the same FI throughout the S&I process.) The FI experience variable used in this dataset reflects experience from screenings.
- **Selected Person-Level File**: A selected person-level file was created for Models 3 and 4 that includes every selected person from 3 survey years: 1999, 2000, and 2001. All the variables that were in the household file also are available in the selected person-level file. In addition, selected person characteristics were available. Even though this dataset represents selected persons, the FI experience variable in this dataset was derived from the screening information. Because it was decided that the first interviewer would be assigned to the whole S&I process for a household, experience at the selected person level and household level is basically identical.
- **Respondent-Level File**: A respondent-level file was created for Model 5 that includes every complete respondent for the 3 survey years: 1999, 2000, and 2001. Because this dataset contains information from the questionnaire, there are a greater number of available variables including substance use characteristics. For this dataset, the interviewer experience variable reflects the actual number of interviews completed instead of screenings completed.

[3] In Quarters 1 and 2 of 2001, an incentive experiment (IE) was conducted within NSDUH. A subset of FI regions was selected into this experiment, and FI regions with historically lower response rates were oversampled. Each FI region has two segments selected per quarter, and for the IE one segment received a control ($0) and one segment received either a $20 or a $40 incentive. As expected, the incentive groups had significantly higher response rates. Therefore, it was important to distinguish effects of this experiment from FI effects in the models. Results are taken from OAS (2002c).

Results

Model 1: Probability of Contacting Household During Screening

The first step in the screening process is to make contact with the household. Using the selected household dataset, household contact was determined by the screening codes in ***Table 5.5*** (codes corresponding to not contacting a household are 11, 12, and 21; otherwise, household contact was made). Household contact is conditional on eligibility. Based on this definition, household contact was modeled using a weighted logistic regression procedure in SUDAAN (Research Triangle Institute [RTI], 2001). Design-based household survey weights also were used. ***Table 5.6*** shows the results of this model. Seeing the effect of FI experience on household contact while controlling for other variables is the main interest (***Figure 5.5***). Highly experienced interviewers had significantly higher odds of contacting a household than did inexperienced interviewers (odds ratio [OR] = 1.54). Interviewers with experience (120-299) had slightly higher although nonsignificant odds of contacting a household. As expected, other FI characteristics did not significantly predict household contact (FI race/ethnicity, gender, or age). In addition, it was hypothesized that a type of "race/ethnicity matching" was occurring where interviewers of different races/ethnicities had different levels of success in certain racial/ethnic concentrations. To account for this phenomenon, interactions of both FI race/ethnicity by Hispanic concentration and FI race/ethnicity by non-Hispanic black concentration were included in the model. However, due to convergence problems, the model was reduced and the interaction of FI race/ethnicity by non-Hispanic black was removed because it provided less information than the interaction of FI race/ethnicity by Hispanic concentration. All the segment-level characteristics except non-Hispanic black concentration were shown to be significant predictors of household contact.

Model 2: Probability of Successful Screening

After contacting the household, the next step in the screening process is to complete a successful screening (referred to as "screening cooperation" in ***Figure 5.6***). Again, using the selected household dataset, successful screening was determined using the codes in ***Table 5.5***. Final screening codes corresponding to a successful screening (household cooperation) are 30, 31, and 32; otherwise, the screening was unsuccessful. Successful screening is conditional on eligibility and contacting the household. That is, Model 2 is conditional on Model 1 (***Figure 5.6***). Based on the screening codes used to represent a successful screening, household cooperation was modeled using a weighted logistic regression procedure in SUDAAN (RTI, 2001). Design-based household survey weights were used, and these weights were adjusted by the predicted propensity taken from Model 1 in order to account for the first stage of the interviewing process. More specifically, the household weight was divided by the predicted propensity taken from Model 1. ***Table 5.7*** shows the results of Model 2. As with all the models, the main interest is to see the effect of FI experience on screening cooperation while controlling for the other variables. Experienced and highly experienced interviewers had increasingly higher odds of successful screening at a household (OR = 1.16 and OR = 1.43, respectively) than inexperienced interviewers. All the segment-level characteristics (and the interaction of FI race/ethnicity and

Table 5.5 Final Screening and Interviewing Codes

Code	Unweighted Frequency	Eligibility (E or I)	Definition
SCREENING			
Household Ineligible			
10	51,319	I	Vacant
13	13,728	I	Not primary residence
20	33	I	Housing unit (HU) listed as group quarters unit (GQU)
22	1,223	I	Contains only military personnel
25	698	I	Other, group quarter is found to be institutionalized, etc.
26	21,451	I	Live there less than ½ quarter (eligibility rule)
29	2,171	I	Listing error
No Household Contact			
11	10,912	E	No one at home after repeated visits
12	1,458	E	Screening respondent unavailable—repeated visits
21	2,706	E	Denied access (building/complex)
Household Contact, but Unsuccessful Screening			
14	1,086	E	Physically/mentally incompetent
15	315	E	Language barrier—Hispanic
16	1,510	E	Language barrier—other
17	27,157	E	Refusal
24	48	E	Other, eligible
33 +	125	E	Problem cases
Successful Screening			
30	292,072	E	No one selected for interview
31	135,018	E	One selected for interview
32	69,322	E	Two selected for interview
INTERVIEW			
Selected Person Ineligible			
81	1,343	I	Other, ineligible selection
82	310	E/I	Used for cases sampled out
83	287	I	Other, 11 year old
84	3	I	Discarded interview (too many selections made)
87	19	I	Other, person in military
Selected Person Not Contacted			
71	5,299	E	No one at dwelling unit (DU) after repeated visits
72	9,908	E	Respondent unavailable after repeated visits
89	74	E	Other, access to building denied (e.g., dorm, complex)
Selected Person Contacted, but Unsuccessful Interview			
74	3,090	E	Physically/mentally incompetent
75	467	E	Language barrier—Hispanic
76	1,391	E	Language barrier—other
77	32,346	E	Refusal
78	8,060	E	Parental refusal for 12 to 17 year old
80	2,038	E	Other, eligible person moved
85	1	E	Questionnaire not returned from field
86	446	E	Eligible person from household (HH) roster who did not return all/part of questionnaire
88	32	E	Other, too dangerous to interview
90 + (except 93)	329	E	Problem cases
Successful Interview			
70	207,785	E	Interview complete
73	401	E	Breakoff (partial interview)
93	6	E	Completed interview, verification revealed that interviewer did not follow correct procedures

E = eligible; I = ineligible.

Table 5.6 Model 1—Probability of Contacting Household During Screening, Given That Household Is Eligible

Variable	Beta	*P* Value (Beta)	Odds Ratio	Wald *F* Statistic	*P* Value (Wald *F*)
Intercept	4.02	0.0000	55.61+	N/A	.
FI Experience				37.41	0.0000
Inexperienced (0-119 Screenings) (RC)	0.00	.	1.00		
Experienced (120-299)	0.09	0.2131	1.10		
Highly Experienced (300+)	0.43	0.0000	1.54+		
FI Race/Ethnicity				N/A	.
White (RC)	0.00	.	1.00		
Black	-0.17	0.5830	0.85		
Hispanic	0.26	0.1792	1.30		
Other	0.08	0.7784	1.08		
FI Gender				1.23	0.2668
Male (RC)	0.00	.	1.00		
Female	0.07	0.2668	1.07		
FI Age				1.04	0.3849
0-40 (RC)	0.00	.	1.00		
41-50	-0.05	0.5030	0.95		
51-60	-0.07	0.2255	0.93		
61+	0.07	0.3871	1.07		
Missing	0.00	1.0000	1.00		
Owner-Occupied Households				42.65	0.0000
≥50% (RC)	0.00	.	1.00		
10%-49%	-0.53	0.0000	0.59+		
<10%	-0.80	0.0000	0.45+		
Population Density				33.65	0.0000
≥1 Million (MSA)	-0.45	0.0000	0.64+		
<1 Million (MSA)	-0.15	0.0156	0.86+		
Non-MSA (RC)	0.00	.	1.00		
Census Region				8.24	0.0000
Northeast	-0.28	0.0100	0.76+		
Midwest	-0.12	0.2700	0.88		
South	0.02	0.8600	1.02		
West (RC)	0.00	.	1.00		
Non-Hispanic Black Concentration				1.15	0.3167
≥50% (RC)	0.00	.	1.00		
10%-49%	0.04	0.7285	1.04		
<10%	0.14	0.1872	1.15		
Hispanic Concentration				N/A	.
≥71% (RC)	0.00	.	1.00		
20%-70%	-0.04	0.7834	0.96		
<20%	-0.31	0.0388	0.73+		

(continued)

Table 5.6 Model 1—Probability of Contacting Household During Screening, Given That Household Is Eligible (continued)

Variable	Beta	*P* Value (Beta)	Odds Ratio	Wald *F* Statistic	*P* Value (Wald *F*)
FI Race/Ethnicity*Hispanic				1.91	0.0764
White, ≥71% (RC)	0.00	.	1.00		
White, 20%-70% (RC)	0.00	.	1.00		
White, <20% (RC)	0.00	.	1.00		
Black, ≥71% (RC)	0.00	.	1.00		
Black, 20%-70%	0.00	0.9957	1.00		
Black, <20%	-0.16	0.6201	0.85		
Hispanic, ≥71% (RC)	0.00	.	1.00		
Hispanic, 20%-70%	-0.80	0.0044	0.45+		
Hispanic, <20%	-0.60	0.0077	0.55+		
Other, ≥71% (RC)	0.00	.	1.00		
Other, 20%-70%	-0.09	0.8020	0.91		
Other, <20%	-0.19	0.5235	0.83		

FI = field interviewer; MSA = metropolitan statistical area; N/A = not available; RC = reference category.
+ Significant at 0.05.

Figure 5.5 Process Measured in Statistical Model 1

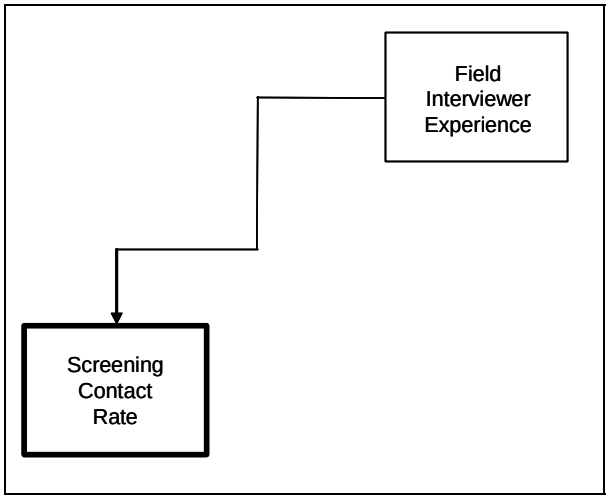
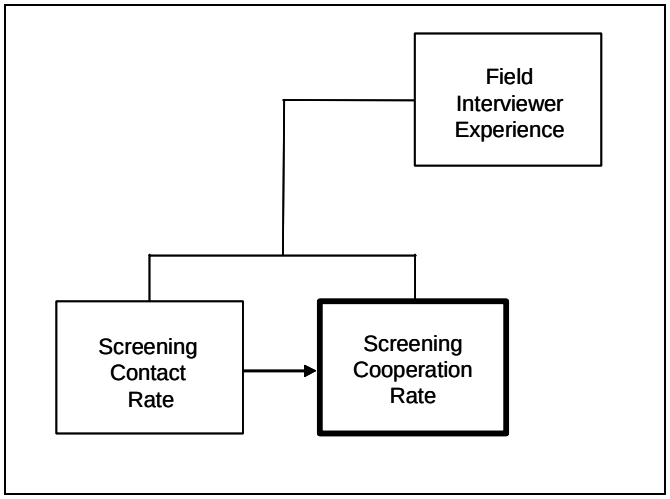


Figure 5.6 Process Measured in Statistical Model 2



Hispanic concentration[4]) and FI characteristics were shown to significantly predict a successful screening. Also, the incentive experiment variable was added to Model 2 because it was felt that at this stage incentives could influence the ability of an FI to obtain a successful screening. As expected, the non-incentive FI regions had higher odds of successful screenings (OR = 1.18) due to the nature of the incentive experiment, which oversampled FI regions with low response rates. However, it was unexpected that within the incentive experiment, FI regions receiving incentives had a lower odds of successful screening (OR = 0.79). This result was verified in the raw data.

[4] For comparability with Model 1, Model 2 included only the interaction of FI race/ethnicity by Hispanic concentration in order to account for "race/ethnicity matching."

Table 5.7 Model 2—Probability of Successful Screening, Given Eligibility and Contact

Variable	Beta	*P* Value (Beta)	Odds Ratio	Wald *F* Statistic	*P* Value (Wald *F*)
Intercept	3.78	0.0000	43.99+	N/A	.
FI Experience				139.16	0.0000
Inexperienced (0-119 Screenings) (RC)	0.00	.	1.00		
Experienced (120-299)	0.15	0.0000	1.16+		
Highly Experienced (300+)	0.36	0.0000	1.43+		
FI Race/Ethnicity				N/A	.
White (RC)	0.00	.	1.00		
Black	-0.50	0.0829	0.61		
Hispanic	0.16	0.3924	1.17		
Other	-0.02	0.9296	0.98		
FI Gender				49.21	0.0000
Male (RC)	0.00	.	1.00		
Female	0.16	0.0000	1.18+		
FI Age				2.95	0.0196
0-40 (RC)	0.00	.	1.00		
41-50	-0.05	0.1126	0.95		
51-60	-0.06	0.0167	0.94+		
61+	0.02	0.5239	1.02		
Missing	-0.03	0.3745	0.97		
Owner-Occupied Households				4.78	0.0086
≥50% (RC)	0.00	.	1.00		
10%-49%	-0.09	0.0070	0.92+		
<10%	-0.11	0.0253	0.90+		
Population Density				283.72	0.0000
≥1 Million (MSA)	-0.62	0.0000	0.54+		
<1 Million (MSA)	-0.35	0.0000	0.71+		
Non-MSA (RC)	0.00	.	1.00		
Census Region				33.51	0.0000
Northeast	-0.13	0.0000	0.88+		
Midwest	0.03	0.3400	1.03		
South	0.16	0.0000	1.17+		
West (RC)	0.00	.	1.00		
Non-Hispanic Black Concentration				20.97	0.0000
≥50% (RC)	0.00	.	1.00		
10%-49%	-0.23	0.0000	0.79+		
<10%	-0.29	0.0000	0.75+		
Hispanic Concentration				N/A	.
≥71% (RC)	0.00	.	1.00		
20%-70%	-0.61	0.0003	0.54+		
<20%	-0.77	0.0000	0.46+		
Incentive Group				44.58	0.0000
Not in Incentive FI Region	0.16	0.0000	1.18+		
Incentive FI Region but Received No Incentive (RC)	0.00	.	1.00		
Incentive FI Region and Received Incentive	-0.23	0.0015	0.79+		

(continued)

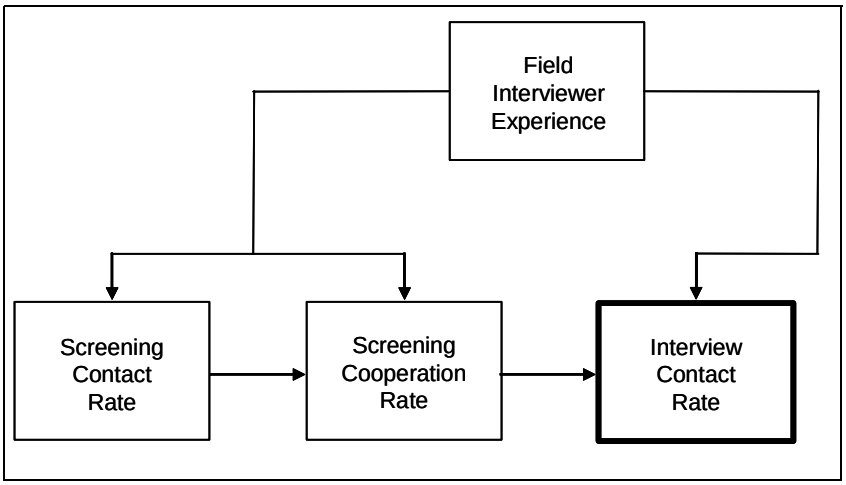
Table 5.7 Model 2—Probability of Successful Screening Given Eligibility and Contact (continued)

Variable	Beta	*P* Value (Beta)	Odds Ratio	Wald *F* Statistic	*P* Value (Wald *F*)
FI Race/Ethnicity*Hispanic				3.22	0.0039
White, ≥71% (RC)	0.00	.	1.00		
White, 20%-70% (RC)	0.00	.	1.00		
White,<20% (RC)	0.00	.	1.00		
Black, ≥71% (RC)	0.00	.	1.00		
Black, 20%-70%	0.65	0.0290	1.92+		
Black,<20%	0.47	0.1012	1.61		
Hispanic, ≥71% (RC)	0.00	.	1.00		
Hispanic, 20%-70%	-0.23	0.2636	0.79		
Hispanic,<20%	-0.36	0.0601	0.69		
Other, ≥71% (RC)	0.00	.	1.00		
Other, 20%-70%	-0.01	0.9739	0.99		
Other,<20%	-0.06	0.7997	0.94		

FI = field interviewer; MSA = metropolitan statistical area; N/A = not available; RC = reference category.
+ Significant at 0.05.

Model 3: Probability of Contacting Selected Person in Household

Figure 5.7 Process Measured in Statistical Model 3



After a successful screening, the next step in the interviewing process is to contact the selected person(s) in the household (referred to as "interview contact" in ***Figure 5.7***). Unlike Models 1 and 2, Model 3 uses the selected person dataset, where selected person contact was determined using the final interview codes shown in ***Table 5.5***. Interview codes corresponding to not contacting selected persons include 71, 72, and 89; otherwise, contact was made. Contacting the selected person is conditional on household eligibility, household contact, and successful screening. That is, Model 3 is conditional on Models 1 and 2 (***Figure 5.7***). Based on this definition, selected person contact was modeled using a weighted logistic regression procedure in SUDAAN using design-based person survey weights. Similar to Model 2, the person weight was divided by the predicted propensity taken from Models 1 and 2 in order to account for the first two stages of the interviewing process. ***Table 5.8*** shows the result of this model. As with all the models, the main interest is to see the effect of FI experience on contacting the selected person while controlling for the other variables. Highly experienced interviewers had significantly higher odds of contacting the selected person (OR = 1.14) than inexperienced interviewers. Other covariates in

Table 5.8 Model 3—Probability of Contacting Selected Person (Interview Stage) in Household, Given Models 1 and 2

Variable	Beta	*P* Value (Beta)	Odds Ratio	Wald *F* Statistic	*P Value* (Wald *F*)
Intercept	3.60	0.0000	36.67+	N/A	.
FI Experience				10.11	0.0000
Inexperienced (0-119 Screenings) (RC)	0.00	.	1.00		
Experienced (120-299)	-0.04	0.4152	0.96		
Highly Experienced (300+)	0.13	0.0005	1.14+		
FI Race/Ethnicity				N/A	.
White (RC)	0.00	.	1.00		
Black	-0.35	0.0000	0.71+		
Hispanic	-0.50	0.0000	0.61+		
Other	-0.09	0.4083	0.91		
FI Gender				0.01	0.9363
Male (RC)	0.00	.	1.00		
Female	0.00	0.9363	1.00		
FI Age				0.67	0.6122
0-40 (RC)	0.00	.	1.00		
41-50	0.03	0.5327	1.03		
51-60	0.00	0.9841	1.00		
61+	0.08	0.1848	1.08		
Missing	0.03	0.6932	1.03		
Owner-Occupied Households				1.37	0.2549
≥50% (RC)	0.00	.	1.00		
10%-49%	-0.07	0.1128	0.93		
<10%	-0.06	0.4006	0.94		
Population Density				23.29	0.0000
≥1 Million (MSA)	-0.24	0.0000	0.79+		
<1 Million (MSA)	-0.04	0.3730	0.96		
Non-MSA (RC)	0.00	.	1.00		
Census Region				27.83	0.0000
Northeast	-0.43	0.0000	0.65+		
Midwest	-0.17	0.0000	0.85+		
South	-0.22	0.0000	0.80+		
West (RC)	0.00	.	1.00		
Selected Person Race/Ethnicity				N/A	.
White (RC)	0.00	.	1.00		
Black	-0.48	0.0000	0.62+		
Hispanic	-0.21	0.0020	0.81+		
Other	-0.12	0.1999	0.89		
Selected Person Gender				52.91	0.0000
Male (RC)	0.00	.	1.00		
Female	0.22	0.0000	1.25+		
Selected Person Age Group				314.74	0.0000
12-17 (RC)	0.00	.	1.00		
18-25	-0.98	0.0000	0.37+		
26-34	-0.93	0.0000	0.40+		
35-49	-0.72	0.0000	0.48+		
50+	-0.28	0.0000	0.76+		
Incentive Group				13.94	0.0000
Not in Incentive FI Region	0.11	0.0027	1.12+		
Incentive FI Region but Received No Incentive (RC)	0.00	.	1.00		
Incentive FI Region and Received Incentive	0.68	0.0000	1.98+		

(continued)

Table 5.8 Model 3—Probability of Contacting Selected Person (Interview Stage) in Household, Given Models 1 and 2 (continued)

Variable	Beta	*P* Value (Beta)	Odds Ratio	Wald *F* Statistic	*P* Value (Wald *F*)
Number of Persons Selected				3.73	0.0539
One person selected (RC)	0.00	.	1.00		
Two persons selected	-0.06	0.0539	0.94		
FI Race/Ethnicity*Selected Person Race/Ethnicity				4.36	0.0000
White, White (RC)	0.00	.	1.00		
White, Black (RC)	0.00	.	1.00		
White, Hispanic (RC)	0.00	.	1.00		
White, Other (RC)	0.00	.	1.00		
Black, White (RC)	0.00	.	1.00		
Black, Black	0.47	0.0000	1.60+		
Black, Hispanic	0.25	0.1000	1.29		
Black, Other	0.37	0.0538	1.45		
Hispanic, White (RC)	0.00	.	1.00		
Hispanic, Black	0.56	0.0006	1.76+		
Hispanic, Hispanic	0.44	0.0045	1.56+		
Hispanic, Other	0.37	0.1272	1.45		
Other, White (RC)	0.00	.	1.00		
Other, Black	0.42	0.1499	1.52		
Other, Hispanic	0.36	0.1186	1.43		
Other, Other	0.06	0.8073	1.06		

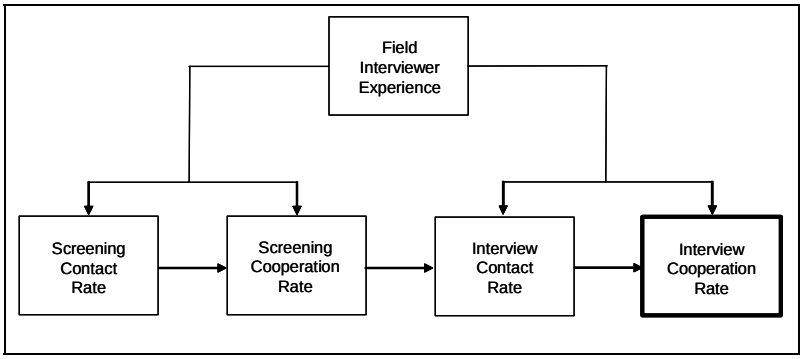
FI = field interviewer; MSA = metropolitan statistical area; N/A = not available; RC = reference category.
+ Significant at 0.05.

the model also were shown to have a significant effect on contacting the selected person. These include population density, census region, incentive group, number of persons selected in the household, the selected person's age and gender, and the interaction of FI race/ethnicity and selected person's race/ethnicity.

Model 4: Probability of Successfully Interviewing Selected Person

After contacting the selected person within a household, the next step in the interviewing process is to successfully interview the selected person (i.e., the selected person completes the questionnaire; in ***Figure 5.8***, this is referred to as "interview cooperation"). Similar to Model 3, Model 4 uses the selected person dataset, where a successful interview was determined using the final interview codes shown in ***Table 5.5***. Final codes corresponding to a successful interview are 70, 73, and 93; otherwise, the interview was not complete. In addition to using the interviewing codes, a usable case rule was used when determining if an interview was successful. A respondent was required to answer a certain number of core questions in order for the interview

Figure 5.8 Process Measured in Statistical Model 4



to be considered complete.[5] A successful interview is conditional on household eligibility, household contact, successful screening, and contacting the selected person. More specifically, Model 4 is conditional on Models 1 through 3 (***Figure 5.8***). Based on this definition, a successful interview was modeled using a weighted logistic regression procedure in SUDAAN. In Model 4, the weight used was the Model 3 weight divided by the Model 3 predicted propensity. ***Table 5.9*** shows the result of this model. As with all the models, the main interest is to see the effect of FI experience on contacting the selected person while controlling for the other variables. Experienced and highly experienced interviewers had increasingly higher odds of obtaining a successful interview (OR = 1.11 and OR = 1.33, respectively) than inexperienced interviewers. In fact, every independent variable in Model 4, except for FI gender and percentage owner-occupied households, predicted significant differences in interview rates.

Effect of Interviewer Experience on Overall Response Rates. Assuming independence of the four rates (screening contact, screening completion, interview contact, and interview completion), an overall response rate can be computed as the product of the four rates. Models 1 through 4 used logistic regression to adjust for other covariates that were expected to influence response rates and estimated the impact of interviewer experience at the three defined levels (inexperienced, experienced, and highly experienced) after adjusting for these covariates. The logistic regression, however, provided estimates of the relationship among the interviewer experience effects in terms of the estimated beta coefficients of the logistic regression model or in terms of estimated odds ratios, but did not provide estimates of the level of the response rate by interview experience. For each model, weighted estimates of the population size associated with each interviewer experience level were estimated by summing the adjusted weights for eligible sample dwelling units (Model 1), for contacted eligible sample dwelling units (Model 2), for eligible sample persons (Model 3), or for contacted eligible sample persons (Model 4). The level of the adjusted response rates by interviewer experience categories was set by requiring that their weighted average equal the unadjusted average for the whole population. Symbolically, this additional constraint can be written as

$$\frac{\sum_{i=1}^{3} w_i r_i^{adj}}{\sum_{i=1}^{3} w_i} = \bar{r},$$

where w_i is the estimated population size for interviewer experience level i, r_i^{adj} is the adjusted rate for interviewer experience level i, and $\bar{r}$ is the unadjusted rate over all three levels of interviewer experience.

[5] The NSDUH usable case rule requires that the lifetime cigarette question and 9 out of the 13 remaining lifetime drug use questions be answered.

Table 5.9 Model 4—Probability of Successfully Interviewing Selected Person, Given Models 1 to 3

Variable	Beta	*P* Value (Beta)	Odds Ratio	Wald *F* Statistic	*P* Value (Wald *F*)
Intercept	1.44	0.000	4.22+	N/A	.
FI Experience				65.46	0.0000
Inexperienced (0-119 Screenings) (RC)	0.00	.	1.00		
Experienced (120-299)	0.10	0.0017	1.11+		
Highly Experienced (300+)	0.29	0.0000	1.33+		
FI Race/Ethnicity				N/A	.
White (RC)	0.00	.	1.00		
Black	-0.02	0.6262	0.98		
Hispanic	-0.09	0.1108	0.91		
Other	-0.04	0.4679	0.96		
FI Gender				2.04	0.1539
Male (RC)	0.00	.	1.00		
Female	0.04	0.1539	1.04		
FI Age				5.21	0.0004
0-40 (RC)	0.00	.	1.00		
41-50	-0.06	0.0274	0.94+		
51-60	-0.03	0.3054	0.97		
61+	-0.13	0.0001	0.88+		
Missing	0.01	0.7622	1.01		
Owner-Occupied Households				1.89	0.1519
≥50% (RC)	0.00	.	1.00		
10%-49%	0.02	0.4613	1.02		
<10%	0.09	0.0680	1.09		
Population Density				81.73	0.0000
≥1 Million (MSA)	-0.34	0.0000	0.71+		
<1 Million (MSA)	-0.21	0.0000	0.81+		
Non-MSA (RC)	0.00	.	1.00		
Census Region				11.89	0.0000
Northeast	0.02	0.6011	1.02		
Midwest	0.07	0.0709	1.07+		
South	0.16	0.0000	1.17+		
West (RC)	0.00	.	1.00		
Selected Person Race/Ethnicity				N/A	.
White (RC)	0.00	.	1.00		
Black	0.12	0.0053	1.13+		
Hispanic	0.34	0.0000	1.41+		
Other	-0.26	0.0000	0.77+		
Selected Person Gender				40.75	0.0000
Male (RC)	0.00	.	1.00		
Female	0.12	0.0000	1.12+		
Selected Person Age Group				316.34	0.0000
12-17 (RC)	0.00	.	1.00		
18-25	-0.17	0.0000	0.85+		
26-34	-0.32	0.0000	0.73+		
35-49	-0.46	0.0000	0.63+		
50+	-0.46	0.0000	0.47+		
Incentive Group				28.21	0.0000
Not in Incentive FI Region	0.13	0.0000	1.14+		
Incentive FI Region but Received No Incentive (RC)	0.00	.	1.00		
Incentive FI Region and Received Incentive	0.42	0.0000	1.52+		
Number of Persons Selected				5.65	0.0177
One person selected (RC)	0.00	.	1.00		
Two persons selected	-0.05	0.0177	0.95+		

(continued)

Table 5.9 Model 4—Probability of Successfully Interviewing Selected Person, Given Models 1 to 3 (continued)

Variable	Beta	*P* Value (Beta)	Odds Ratio	Wald *F* Statistic	*P* Value (Wald *F*)
FI Race/Ethnicity*Selected Person Race/Ethnicity				4.30	0.0000
White, White (RC)	0.00	.	1.00		
White, Black (RC)	0.00	.	1.00		
White, Hispanic (RC)	0.00	.	1.00		
White, Other (RC)	0.00	.	1.00		
Black, White (RC)	0.00	.	1.00		
Black, Black	0.34	0.0000	1.41+		
Black, Hispanic	0.31	0.0412	1.37+		
Black, Other	0.19	0.2414	1.21		
Hispanic, White (RC)	0.00	.	1.00		
Hispanic, Black	0.27	0.0570	1.31		
Hispanic, Hispanic	0.26	0.0070	1.29+		
Hispanic, Other	-0.45	0.0061	0.64+		
Other, White (RC)	0.00	.	1.00		
Other, Black	0.01	0.9531	1.01		
Other, Hispanic	0.12	0.3509	1.13		
Other, Other	-0.04	0.8415	0.96		

FI = field interviewer; MSA = metropolitan statistical area; N/A = not available; RC = reference category.
+ Significant at 0.05.

With this additional constraint, the adjusted rate for interviewer experience level 1 was first set arbitrarily and the odds ratios for levels 2 and 3 were used to obtain the adjusted rates for levels 2 and 3. The weighted sum then was compared with the overall unadjusted rate. Iterative interpolation was used to adjust the level 1 rate and recompute the level 2 and 3 rates until the constraint above was satisfied.

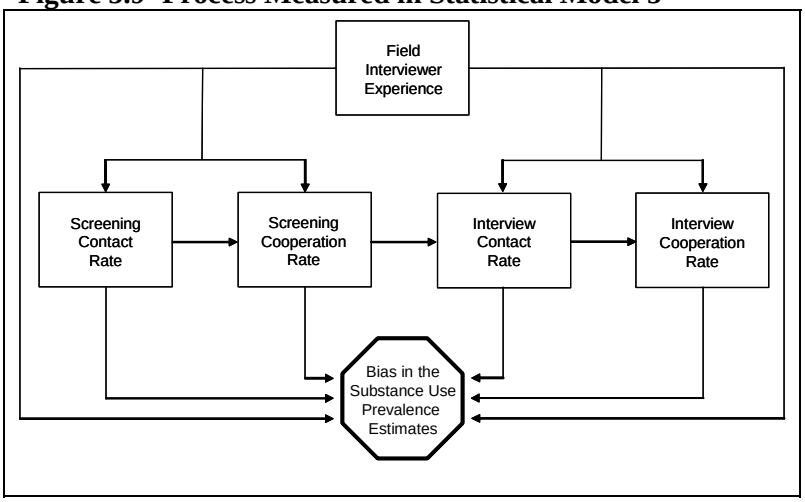
Table 5.10 shows the results of these computations and the products of the four rates as an overall response rate. Inexperienced interviews achieved an (adjusted) response rate of 61.1 percent. Experienced interviewers achieved an adjusted response rate of 63.5 percent, an increase of 2.4 percent over the inexperienced interviewers. Highly experienced interviewers achieved an (adjusted) response rate of 68.4 percent, an increase of 4.9 percent over experienced interviewers and 7.3 percent over inexperienced interviewers.

Table 5.10 Adjusted Response Rate at Each Interview Stage, by Interviewer Experience

No.	Interview Stage	Interviewer Experience		
		Inexperienced	Experienced	Highly Experienced
1	Contacting Household	0.9629	0.9661	0.9756
2	Gaining Household Cooperation/Successful Screening	0.9285	0.9378	0.9489
3	Contacting Selected Person	0.9372	0.9348	0.9445
4	Interviewing Selected Person	0.7296	0.7497	0.7821
	Overall Response Rate	0.6113	0.6349	0.6838

Model 5: Probability of Respondent Reporting Lifetime Substance Use

Figure 5.9 Process Measured in Statistical Model 5



After a selected person responds to the interview, the next step in the interviewing process is to measure substance use. Model 5 uses the respondent person dataset. In this case, reports of substance use are the variables of interest; they are categorized by lifetime use, past year use, and past month use. Model 5 is conditional on each of the four previous models (***Figure 5.9***). The person weight used was taken from Model 4 and then adjusted by dividing the weight by the response propensity of the fourth model. Models were run on the following substance use measures: lifetime, past year, and past month any illicit drug use; lifetime, past year, and past month marijuana use; lifetime, past year, and past month any nonmedical psychotherapeutic use; lifetime, past year, and past month nonmedical pain reliever use; lifetime, past year, and past month alcohol use; and lifetime, past year, and past month cigarette use. Although a greater number of covariates were available with this model, it was felt that only covariates that could not be influenced by FI experience should remain in the model. More specifically, variables such as short interview time or reporting use of other substances were classified as possible confounders of FI experience and thus were excluded from the model. However, the model contains known sociodemographic correlates of substance use based on previous NSDUH experience with the addition of FI characteristics and FI experience. Each model used the same set of covariates, and the only differences were whether variables were taken out due to convergence problems.

Due to the large number of models, only the detailed results for lifetime any illicit drug use will be displayed. Any illicit drug use includes marijuana/hashish, cocaine (including crack), heroin, hallucinogens, inhalants, or any prescription-type psychotherapeutic used nonmedically. ***Table 5.11*** shows the results of this model. As with all the models, the main interest is to see the effect of FI experience on drug use prevalence rates while controlling for the other variables. Experienced and highly experienced interviewers had increasingly lower odds of a respondent's reporting any illicit lifetime usage (OR = 0.91 and OR = 0.83, respectively) than inexperienced interviewers. In fact, every independent variable in this model significantly predicted any illicit lifetime use as compared with the reference cell except for FI gender (p = 0.68). Unfortunately, one variable, FI age category, was taken out of the model due to convergence problems.

Table 5.11 Model 5—Probability of Respondent Reporting Lifetime Any Illicit Drug Use, Given Models 1 to 4 with Propensity Adjustment

Variable	Beta	*P* Value (Beta)	Odds Ratio	Wald *F* Statistic	*P* Value (Wald *F*)
Intercept	-0.81	0.000	0.44+	N/A	.
FI Experience				25.81	0.0000
Inexperienced (0-39 Interviews) (RC)	0.00	.	1.00		
Experienced (40-99)	-0.09	0.0003	0.91+		
Highly Experienced (100+)	-0.19	0.0000	0.83+		
FI Race/Ethnicity				16.54	0.0000
White (RC)	0.00	.	1.00		
Black	-0.15	0.0000	0.86+		
Hispanic	-0.28	0.0000	0.76+		
Other	0.02	0.766	1.02		
FI Gender				0.17	0.6796
Male (RC)	0.00	.	1.00		
Female	-0.01	0.6777	0.99		
Total Family Income				178.61	0.0000
<$20,000	0.29	0.0000	1.34+		
≥$20,000 (RC)	0.00	.	1.00		
Population Density				75.62	0.0000
≥1 million (MSA)	0.29	0.0000	1.33+		
<1 million (MSA)	0.24	0.0000	1.28+		
Non-MSA (RC)	0.00	.	1.00		
Census Region				58.97	0.0000
Northeast	-0.28	0.0000	0.75+		
Midwest	-0.32	0.0000	0.73+		
South	-0.36	0.0000	0.70+		
West (RC)	0.00	.	1.00		
Respondent Race/Ethnicity				188.87	0.0000
White (RC)	0.00	.	1.00		
Black	-0.25	0.0000	0.78+		
Hispanic	-0.67	0.0000	0.51+		
Other	-0.99	0.0000	0.37+		
Respondent Gender				263.77	0.0000
Male	0.00	.	1.00		
Female	-0.28	0.0000	0.75+		
Respondent Age Group				729.86	0.0000
12-17 (RC)	0.00	.	1.00		
18-25	0.78	0.0000	2.18+		
26-34	1.00	0.0000	2.73+		
35+	0.41	0.0000	1.51+		
Respondent Marital Status					
Married	0.00	.	1.00	923.82	0.0000
Widowed	-1.21	0.0000	0.30+		
Divorced/Separated	0.65	0.0000	1.92+		
Never Married	0.49	0.0000	1.63+		
≤14 Years Old	-0.70	0.0000	0.49+		
Incentive Group				5.14	0.0061
Not in Incentive FI Region	-0.03	0.1315	0.97		
Incentive FI Region but Received No Incentive (RC)	0.00	.	1.00		
Incentive FI Region and Received Incentive	0.14	0.0303	1.15+		

(continued)

Table 5.11 Model 5—Probability of Respondent Reporting Lifetime Any Illicit Drug Use, Given Models 1 to 4 with Propensity Adjustment (continued)

Variable	Beta	*P* Value (Beta)	Odds Ratio	Wald *F* Statistic	*P* Value (Wald *F*)
Number of Respondents in HH				420.69	0.0000
One Person in HH (RC)	0.00	.	1.00		
Two Persons in HH	0.38	0.0000	1.46+		
Survey Year				17.74	0.0000
1999	-0.13	0.0000	0.87+		
2000	-0.12	0.0000	0.88+		
2001 (RC)	0.00	.	1.00		

FI = field interviewer; HH = household; MSA = metropolitan statistical area; N/A = not available; RC = reference category.
+ Significant at 0.05.

A major focus of this research was to attempt to understand the relationship between response rates and prevalence rates as interviewer experience increases and to be able to test hypotheses about this relationship. Model-adjusted rates were developed based on Model 5 using the predictive marginal means option of SUDAAN's PROC MULTILOG. Predictive means represent the prevalence rate that would have occurred if all interviewers had the specified level of experience (Korn & Graubard, 1999). The predictive marginal means are shown as the adjusted rates shown in ***Tables 5.12, 5.13***, and ***5.14***. Note that in every case the adjusted rate estimates decrease as interviewer experience increases from inexperienced to experienced to highly experienced.

Table 5.12 Adjusted and Incremental *Lifetime* Prevalence Rates, by Interviewer Experience

Lifetime Prevalence Rate	Interviewer Experience		
	Inexperienced	Experienced	Highly Experienced
Any Illicit Drug			
Adjusted Rate	0.422	0.402[b]	0.382[b]
Incremental Rate		-0.134	0.122
Marijuana			
Adjusted Rate	0.363	0.352[a]	0.339[a]
Incremental Rate		0.052	0.178
Any Nonmedical Psychotherapeutics			
Adjusted Rate	0.170	0.155[b]	0.139[b]
Incremental Rate		-0.243[c]	-0.073
Nonmedical Pain Relievers			
Adjusted Rate	0.103	0.092[b]	0.082[b]
Incremental Rate		-0.212[d]	-0.047
Alcohol			
Adjusted Rate	0.821	0.814	0.807
Incremental Rate		0.635	0.711
Cigarettes			
Adjusted Rate	0.685	0.666[b]	0.664
Incremental Rate		0.194	0.636

[a] Difference between adjusted rates in this column and previous column is statistically significant at 0.05 level.
[b] Difference between adjusted rates in this column and previous column is statistically significant at 0.01 level.
[c] Incremental rate is significantly negative at 0.05 level.
[d] Incremental rate is significantly negative at 0.01 level.

Table 5.13 Adjusted and Incremental *Past Year* Prevalence Rates, by Interviewer Experience

Past Year Prevalence Rate	Interviewer Experience		
	Inexperienced	Experienced	Highly Experienced
Any Illicit Drug			
Adjusted Rate	0.126	0.117[b]	0.108[b]
Incremental Rate		-0.120	-0.008
Marijuana			
Adjusted Rate	0.091	0.086[a]	0.083
Incremental Rate		-0.057	0.045
Any Nonmedical Psychotherapeutics			
Adjusted Rate	0.049	0.045	0.038[b]
Incremental Rate		-0.057	-0.056[d]
Nonmedical Pain Relievers			
Adjusted Rate	0.036	0.034	0.028[b]
Incremental Rate		-0.033	-0.048[d]
Alcohol			
Adjusted Rate	0.637	0.626[a]	0.619
Incremental Rate		0.321	0.528
Cigarettes			
Adjusted Rate	0.299	0.292	0.287
Incremental Rate		0.117	0.214

[a] Difference between adjusted rates in this column and previous column is statistically significant at 0.05 level.
[b] Difference between adjusted rates in this column and previous column is statistically significant at 0.01 level.
[c] Incremental rate is significantly negative at 0.05 level.
[d] Incremental rate is significantly negative at 0.01 level.

Table 5.14 Adjusted and Incremental *Past Month* Prevalence Rates, by Interviewer Experience

Past Month Prevalence Rate	Interviewer Experience		
	Inexperienced	Experienced	Highly Experienced
Any Illicit Drug			
Adjusted Rate	0.070	0.064[a]	0.061
Incremental Rate		-0.086	0.022
Marijuana			
Adjusted Rate	0.052	0.048[a]	0.047
Incremental Rate		-0.053	0.025
Any Nonmedical Psychotherapeutics			
Adjusted Rate	0.021	0.020	0.016[b]
Incremental Rate		-0.015	-0.024
Nonmedical Pain Relievers			
Adjusted Rate	0.015	0.013	0.012
Incremental Rate		-0.038	-0.012
Alcohol			
Adjusted Rate	0.479	0.469	0.465
Incremental Rate		0.191	0.421
Cigarettes			
Adjusted Rate	0.257	0.249	0.245
Incremental Rate		0.033	0.199

[a] Difference between adjusted rates in this column and previous column is statistically significant at 0.05 level.
[b] Difference between adjusted rates in this column and previous column is statistically significant at 0.01 level.
[c] Incremental rate is significantly negative at 0.05 level.
[d] Incremental rate is significantly negative at 0.01 level.

In summary, the impact of increased interviewer experience was to simultaneously increase response rates and decrease estimated prevalence rates. What could explain these observed phenomena? Three possible hypotheses are summarized as follows:

Hypothesis 1: The decrease in estimated substance use can be explained by lower use rates among the additional selected persons who respond to interviewers who have more experience.
Hypothesis 2: Persons interviewed by more experienced interviewers generally report lower substance use regardless of their propensity to respond.
Hypothesis 3: The decrease in prevalence rates associated with more experienced interviewers is explained by some mix of differing substance use among new respondents and differing reporting of substance use among all respondents, resulting in a net decrease in estimated substance use.

If Hypothesis 1 is accepted, then selected persons who respond to inexperienced interviewers would report the same substance use when interviewed by more experienced interviewers. In addition, the hypothesis implies that all respondents at a specified level of interviewing also would be respondents at a higher level of experience. These assumptions were applied to computing incremental prevalence rates for the additional persons who respond to experienced and highly experienced interviewers. Incremental rates for level of experience $i \geq 2$ are defined by the following relationship:

$$R_{i-1}P_{i-1} + (R_i - R_{i-1})I_i = R_iP_i ,$$

where R_i is the response rate at experience level i, P_i is the adjusted prevalence rate at experience level i, and I_i is the incremental prevalence rate at experience level i. The solution for the incremental rate then is expressible as a function of the adjusted rates and the response rates as

$$I_i = \frac{R_i}{R_i - R_{i-1}} P_i - \frac{R_{i-1}}{R_i - R_{i-1}} P_{i-1} \ .$$

Note that $R_i - R_{i-1}$ represents the incremental response rate achieved by interviewers at experience level i compared with interviewers at experience level i-1; from ***Table 5.10***, these incremental response rates were 2.4 percent for increasing inexperience from level 1 to level 2 and 4.9 percent for increasing experience from level 2 to level 3. Because these incremental response rates are small relative to the response rates, the coefficients applied to the adjusted rates in order to compute the incremental rates were relatively large numbers.

Hypothesis 1 would be supported if the incremental prevalence rates at experience levels 2 and 3 could be shown to be lower than the adjusted rates at levels 1 and 2, respectively. Symbolically, it would be best to test and reject the null hypotheses, $H_o : I_i - P_{i-1} = 0$ for $i = 1, 2$. The contrast needed to test this hypothesis can be estimated by $\hat{C}_i = \frac{R_i}{R_i - R_{i-1}} \left(\hat{P}_i - \hat{P}_{i-1} \right)$. If the analysis is made conditional on the response rates (i.e., treating the response rates as fixed or known rather than estimated), then this test is equivalent to testing the null hypothesis of equal

adjusted rates (i.e., $H_o : P_i - P_{i-1} = 0$ for $i = 1, 2$). ***Tables 5.12, 5.13***, and ***5.14*** show the results of applying this test to the marginal rates for the six substance use measures and three levels of recency of use. For lifetime use measures (***Table 5.12***), the hypothesis of no difference in the adjusted rates is rejected when comparing experienced with inexperienced levels for five of the six substance use measures; the hypothesis of no difference is plausible for lifetime alcohol use. When comparing highly experienced with experienced interviewers, the hypothesis of no difference is rejected for four of the six lifetime substance use measures (any illicit drug, marijuana, any nonmedical psychotherapeutic, and nonmedical pain relievers). For past year use measures, the hypothesis of no difference is rejected for 6 of the 12 possible cases, and for past month use measures, the hypothesis of no difference is rejected for only 3 of the tested cases.

From these results, it appears reasonable to assume that for many substance use measures, the incremental prevalence rates associated with the higher response rates attained by more experienced interviewers are indeed lower than the prevalence rates reported by the respondents interviewed by the less experienced interviewers. If Hypothesis 1 were the only explanation for the reduced prevalence rates, then negative incremental rates as computed above are not possible. Because the incremental response rate estimates are computed as a weighted contrast of the adjusted rate estimates and the weights are relatively large numbers, the standard errors of the incremental rates also are large even with the large sample available for these analyses. Hypothesis 2 would be strongly supported if the marginal rates could be shown to be negative (i.e., if the null hypothesis $H_o : I_i \geq 0$ for $i = 1, 2$ could be rejected). This hypothesis was tested for the incremental rates shown in ***Tables 5.12, 5.13,*** and ***5.14***. The hypothesis was rejected for any nonmedical psychotherapeutics and nonmedical pain relievers at the experienced level for lifetime measures shown in ***Table 5.12*** and at the highly experienced level for past year measures in ***Table 5.13***. It could not be rejected for any of the past month substance use measures in ***Table 5.14***.

Even without strong support for Hypothesis 2 implied by the hypothesis tests for positive incremental rates, Hypothesis 2 remains plausible as a contributing factor. Hypothesis 2 could be an explanation for the reduced prevalence rates if the same respondents who would have responded to interviewers with less experience also respond to the more experienced interviewers, but report lower substance use. Hypothesis 2 also could explain the difference in prevalence rates if the more experienced interviewers achieve higher response by substantially increasing the response rates for types of respondents who have low substance use behaviors while slightly decreasing response rates for types of respondents who have higher substance use behaviors.

Neither Hypothesis 1 nor Hypothesis 2 can be shown to clearly explain the simultaneous increase in response rates and decrease in prevalence rates as interviewer experience increases. This admits the plausibility of Hypothesis 3, which incorporates a mix of the effects of the other two hypotheses.

Conclusions

The analysis shows that increased interviewer experience simultaneously increases response rates and decreases prevalence rates. In addition, the effect of increased interviewer experience

on prevalence rates cannot be fully explained by weight adjustments based on earlier models (i.e., screening and interview level). In other words, the interviewer effect on prevalence rates cannot be fully attributed to the increase in response rates by experienced interviewers. Furthermore, interviewer experience was significant in the final model, showing that the covariates also did not account for all the decrease in prevalence rates.

A statistical analysis of marginal and incremental prevalence rates based on three levels of interviewer experience showed that plausible explanations for the decrease in prevalence rates for experienced interviewers include (1) lower substance use reporting by the additional respondents and (2) lower reporting of substance use by respondents that interviewers with all levels of experience interview.

These results have important implications for survey methodology in general and more specifically for NSDUH. It is important to reduce any type of selection bias present in a survey, and maintaining high response rates is key to this goal. If experienced interviewers are obtaining interviews from respondents who are more likely to report lower substance use, this is an important methodological concern. Currently, weighting techniques are used to adjust for these nonrespondents. However, if the nonrespondents are fundamentally different from the respondents and the weights are not able to capture this, bias will be introduced into the estimates. Furthermore, a high standard of interviewing also is key to reducing bias and it is important that the interviewers strictly follow survey protocol. It is believed that the inexperienced interviewers follow interview protocol more closely than experienced interviewers due to the tailoring that the experienced interviewers begin to use after gaining experience on the survey. For example, one tailoring method that breaches survey protocol is interviewers' telling the respondent that saying "no" to drug gate questions makes the interview go faster. This behavior can negatively affect prevalence rates and, as a result, is a threat to the validity of the survey. Ensuring that interviewers strictly follow survey protocol can be achieved through rigorous training and observations throughout the year.

This chapter provides a comprehensive analysis of NSDUH interviewer experience and expands the previous analysis that investigated separately prevalence and response rates in relation to interviewer experience. Yet there are limitations to this work. Interviewer behaviors, which are an important component of the conceptual model, were not accounted for due to inadequate data. FI observations started in 2001 and are ongoing in 2004. These data may be useful in providing anecdotal evidence of interviewer behaviors. One also would like to incorporate verification data, which is another source of obtaining FI behaviors. To use these data, weighting techniques would need to be incorporated because data are collected on a subset of screenings and interviews. Another important improvement to any future analysis would be the creation of a more sensitive measure of experience that incorporates more than the number of screenings or interviews conducted. For instance, it has been suggested that interviewers who conduct a small number of interviews and then terminate employment conduct interviews differently than interviewers who are employed for a substantial amount of time. If this is true, then the early interviews conducted by each group should be analyzed separately and not be combined together in the "inexperienced" group. Another suggestion is to analyze the relationship between prevalence rates and interviewer attrition to determine if the areas with higher substance use are more likely to experience interviewer turnover. Further analysis is planned to verify whether the decrease in reporting substance use occurs among interviewers as

they gain expertise or whether other factors interact with interviewer retention rates to create an artificial association of interviewer experience with substance use.

References

Biemer, P. P. (1988). Measuring data quality. In R. M. Groves, P. P. Biemer, L. Lyberg, J. Massey, W. Nicholls, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 273-282). New York: Wiley.

Biemer, P. P. (2001). Nonresponse bias and measurement bias in a comparison of face to face and telephone interviewing. *Journal of Official Statistics, 17*, 295-320.

Cleary, P. D., Mechanic, D., & Weiss, N. (1981). The effect of interviewer characteristics on responses to a mental health interview. *Journal of Health and Social Behavior, 22*, 183-193.

Eyerman, J., Odom, D., Wu, S., & Butler, D. (2002). Nonresponse in the 1999 NHSDA. In J. Gfroerer, J. Eyerman, & J. Chromy (Eds.), *Redesigning an ongoing national household survey: Methodological issues* (pp. 23-51, DHHS Publication No. SMA 03-3768). Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies. [Available as a PDF at http://www.oas.samhsa.gov/nhsda/methods.cfm#Reports]

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: Wiley.

Hughes, A., Chromy, J., Giacoletti, K., & Odom, D. (2001, August). Impact of interviewer experience on drug use prevalence rates in the 1999 NHSDA. In *Proceedings of the American Statistical Association* [CD-ROM]. Alexandria, VA: American Statistical Association.

Hughes, A., Chromy, J., Giacoletti, K., & Odom, D. (2002). Impact of interviewer experience on respondent reports of substance use. In J. Gfroerer, J. Eyerman, & J. Chromy (Eds.), *Redesigning an ongoing national household survey: Methodological issues* (pp. 161-184, DHHS Publication No. SMA 03-3768). Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies. [Available as a PDF at http://www.oas.samhsa.gov/nhsda/methods.cfm#Reports]

Körmendi, E., & Pedersen, S. (1995). Interviewer effect concerning sensitive questions in surveys. In *Proceedings of the International Conference on Survey Measurement and Process Quality, Bristol, UK* (pp. 139-144). Alexandria, VA: American Statistical Association.

Korn, E. L., & Graubard, B. I. (1999). *Analysis of health surveys.* New York: Wiley.

Office of Applied Studies. (2002a). *Results from the 2001 National Household Survey on Drug Abuse: Volume I. Summary of national findings* (DHHS Publication No. SMA 02-3758, NHSDA Series H-17). Rockville, MD: Substance Abuse and Mental Health Services Administration. [Available at http://www.oas.samhsa.gov/p0000016.htm#Standard]

Office of Applied Studies. (2002b). *Results from the 2001 National Household Survey on Drug Abuse: Volume II. Technical appendices and selected data tables* (DHHS Publication No. SMA 02-3759, NHSDA Series H-18). Rockville, MD: Substance Abuse and Mental Health Services Administration. [Available at http://www.oas.samhsa.gov/p0000016.htm#Standard]

Office of Applied Studies. (2002c). *2001 National Household Survey on Drug Abuse: Incentive experiment combined quarter 1 and quarter 2 analysis*. Rockville, MD: Substance Abuse and Mental Health Services Administration. [Available as a PDF at http://www.oas.samhsa.gov/nhsda/methods.cfm and http://www.oas.samhsa.gov/nhsda/methods/incentive.pdf]

Research Triangle Institute. (2001). *SUDAAN user's manual: Release 8.0.* Research Triangle Park, NC: Author.

Stevens, J. A., & Bailar, B. A. (1976). *The relationship between various interviewer characteristics and the collection of income data*. Paper presented at the annual meeting of the American Statistical Association, Boston, MA.

6. Development of a Spanish Questionnaire for NSDUH

Marjorie Hinsdale, Antonieta Díaz, Christine Salinas, and Jeanne Snodgrass
RTI International

Joel Kennet
Substance Abuse and Mental Health Services Administration

Introduction

Translation of survey questionnaires is becoming standard practice for large-scale data collection efforts as growing numbers of non-English speaking immigrants arrive in the United States. However, the methods used to produce survey translations have not been standardized—even for Spanish, the most common target language (Shin & Bruno, 2003). This chapter describes the techniques and principles that were applied in a multicultural review of the translation of the National Survey on Drug Use and Health (NSDUH) questionnaire.[1] Common problems that arose in the translation process and best practices for their resolution also are discussed. In addition, because increasing numbers of surveys are employing computer-assisted interviewing (CAI), this chapter illustrates some of the ways in which this technology ideally can be put to use in conveying translated materials.

The translation review of NSDUH was carried out for a variety of reasons. For many years, the survey has provided a Spanish-language version of the instrument for respondents who requested it. Each year, as new questions were added to the survey, translations were carried out on an ad hoc basis using a variety of translators. In the 1999 survey redesign, a large number of questions were added, and a large number of existing questions were altered to accommodate the audio computer-assisted self-interviewing (ACASI) format. It became apparent through feedback from the field that some of the Spanish questions seemed awkward, and consequently survey staff decided that a comprehensive review would be appropriate.

It was determined that a multicultural review of the 2000 survey's Spanish instrument would be the most effective procedure. Using three translators coming from varied backgrounds in Central and South America, and focus groups of potential respondents representing the major Spanish-speaking groups in the United States, a translation service that specialized in this type of work carried out a review of the entire questionnaire. The specifics of the focus group and multicultural review processes that took place are described in this chapter within the context of a discussion of "best practices" for the development of Spanish survey translations.

[1] Prior to 2002, the survey was called the National Household Survey on Drug Abuse (NHSDA). In this report, it is referred to as NSDUH, regardless of the survey year.

Several critical steps in the development of accurate Spanish survey translations have been identified. A seemingly obvious first step involves staffing the project with qualified personnel—from translators to interviewers. To optimize respondent comprehension, translations should be developed using a multicultural approach and should be tested and reviewed thoroughly by a diverse group of bilingual persons (Schoua Glusberg, 2000). Understanding and applying the concept of cognitive equivalence versus direct translation are key in the development of an effective survey translation. Just as in questionnaire development of English-language surveys, cognitive testing should be employed to identify and correct potential flaws in wording. For studies such as NSDUH that use ACASI, a professional Spanish-speaking voice and skilled audio technicians are needed to ensure the high quality of the audio recording, which maximizes respondents' comprehension. Bilingual interviewers should be fluent and literate in both Spanish and English, and these skills must be demonstrated using a standardized certification procedure. Finally, allowing sufficient time to implement the Spanish translation and train the interviewing staff is perhaps the most problematic step of all because data collection schedules are typically rigorous and researchers are often challenged to maintain the critical timeline even without translations.

Techniques and Principles for Producing Quality Spanish Translations

Selecting a Translator

The selection of translators must be carefully considered. Characteristics and cultural awareness of the translator(s) and reviewer(s) must be weighed against known attributes and cultural sensitivity of the target population in order to make the best matches. Ethnic origin, length of time in the United States, age, and gender of the sample members all play an important role in the decisions regarding how and what to translate. A translator must be aware of cultural distinctions among varying Spanish-speaking populations that are likely to be present in the survey sample. Ignorance of these distinctions oversimplifies the translation process and risks the integrity of the intent of the question. The translator and the researcher must understand the concepts of cultural equivalence versus direct translation, as this is critical to the development of a quality translation (Behling & Law, 2000). A qualified translator must communicate the capabilities as well as the limitations of the translation so that the researchers are cognizant of the distinctions during the analysis phase.

The translator must be someone who can be trusted to (1) make every effort to preserve the meaning, or cognitive equivalence, of the source questions and instructions in English, and (2) communicate deviations in the target language, Spanish, from the original language when necessary in order to maximize overall comprehension. A poor or inexperienced translator might provide translations that are too literal or he or she may take too many liberties with the interpretation. Although these problems are at opposite ends of the scale, both can be equally problematic for the integrity of the data collected.

Translations that are too loose may sound more natural in the target language, but may bias responses because the intent of the question in English has been adapted to something that "sounds" good, but actually means something different. In cases where a translator takes liberties with a sentence to make it sound better, sometimes the intent of the question is lost or misrepresented. Left uncorrected, this can invalidate the responses to data collection questions

and skew the data. At the other end of the scale, individual words and phrases may be problematic if translated too directly. A skilled translator is capable of identifying these cases and will communicate with the researcher if deviations from the source language are necessary but could be problematic. In general, the translator must understand that, for the most part, the goal is to produce a "pragmatic translation" (Brislin, 1980; Casagrande, 1954), which can be seen as somewhat akin to the translation of a technical instruction manual. The goal is to ensure that respondents understand the intent of the questions, and respond appropriately, in accord with their own knowledge and experience. This is in contrast to a "linguistic translation," wherein literal equivalents of individual words are substituted for the English and arranged within the syntax of the target language.

In many cases, professional translation services can provide the researcher with the expertise required for a translation of high quality. However, when not fully investigated, these services also can wreak havoc with the data collection effort if the quality is not up to the standard of the researcher. Other issues that can be vexing when dealing with translation services include control of work priorities, deadlines, and consistency. Translation services frequently outsource work, so the quality can vary from one project to another and even within the same survey instrument. Depending on the service's experience in dealing with survey researchers, the translators may be less likely to understand the methodology behind the questions and therefore are ill equipped to preserve the meanings. However, it also is true that high-quality translation services exist, and these services can provide excellent translations and feedback to the researcher. Generally speaking, translation services are able to access talent with specific language strengths to match the sample population if the researcher can specify the desired language, ethnic background, or dialect. In the best of cases, the translation service will have particular expertise in social science research in addition to excellent translators. These specialized services, although likely to be somewhat more costly, are preferable to services that only offer translation.

An alternative or adjunct to contracting with translation services is to hire in-house translators. Over time, these translators can be trained in the basics of survey methodology so that they have an understanding of the dangers of subtle modifications, which seem harmless, but can lead to inaccurate data being collected. Another advantage of maintaining the same translators is that they gain experience with each project and can create translation memory tools and a library of relevant information to facilitate future translations. In survey research, similar if not identical phrases frequently are used from one survey to another. In many cases, the same translation may be used if the target population is the same and there are no other extenuating circumstances. In-house translators also are more likely to meet specific project schedules and goals because they do not have competing clients' demands to satisfy. Ultimately, qualified in-house translators may be more reliable than translation services.

Regardless of whether the translator is contracted through a service or is the researcher's employee, the best translators tend to be native speakers of the target language who also are fluent in English. Although it is possible for nonnative speakers to provide accurate translations, the risk of having a translation *sound* translated is increased. Not to oversimplify the issue, native speakers of the Spanish language vary a great deal among themselves. Each Spanish-speaking country has its own idiosyncracies with regard to vocabulary, pronunciation, and norms for informal and formal grammatical sentence structure, to name a few.

In the selection of a translator, the researcher must consider the sample population to be studied and try to match the translator's experience and background accordingly. The orientation of the translator, whether it is Mexican, Puerto Rican, Cuban, a Central or South American country, or Spain, should be matched to the survey population whenever possible. However, national surveys within the United States must attempt to address the needs of many if not all of these cultures. This demonstrates the need for a translation protocol that goes beyond an individual translator working in isolation.

For the translation review carried out in 2000, translators were carefully chosen in such a way that nearly all of the criteria listed above were met. As mentioned earlier, a reputable translation service, with specific experience in Spanish survey translation, was hired. The translation service in turn selected translators who were capable of meeting the demands of this complex task. These individuals came from diverse backgrounds (native-speaking origins in Mexico, Puerto Rico and Peru, respectively), yet were experienced in survey research and sensitive to the idea that cognitive equivalence should take precedence over direct translation. In addition, the translators' work was reviewed and augmented by an in-house, bilingual staff member with high familiarity with the intent of the survey questions, as well as the methods employed in its administration.

Reviewing the Translation

Thus far, the focus has been on selecting appropriate translators and reviewers, but equally important is the selection of an appropriate method of review. There are several options for ensuring the quality of the original translation, including simple comparison, back-translation, and multicultural committee review. As might be expected, the schedule and budget may affect the feasibility of these options. The multicultural committee review process, which was used for the 2000 survey's translation, is by far the most costly of these options, but achieves the best result in translation quality.

At minimum, a translation should be subject to a *simple comparison* of the English and Spanish versions by an independent bilingual translator. In this process, the two translators would meet to discuss any discrepancies and would agree on the best translation. This process helps to identify any careless errors that may have been overlooked and also affords a second opinion for alternative wording in situations that are not clear. This process is not as rigorous as some of the other means of ensuring quality control, and it is subject to the biases of the two translators' dialects.

Another common review process is known as *back translation* (Brislin, 1970, 1980). In this process, the original translator uses the source document to create a Spanish translation. Once complete, the translated document then is given to an independent translator who is instructed to translate the Spanish document into English. Both bilingual and nonbilingual staff then compare the two English documents (original and back-translated) to identify potential problems with the translation. In theory, this process sounds like a useful tool, especially for a nonbilingual researcher who would like to be more involved with the quality control of the translation. However, the reality is that this process of translating and back-translating is limited in its usefulness and is time-consuming when done correctly. Consider the following example from the survey instrument:

> Throughout the core modules, the lead-in question, "Have you ever used [DRUG NAME] even once?" appears. One problem with translating this particular question is that the word "ever" simply does not exist in Spanish. For this reason, translations are necessarily different from the English, and back-translations are not particularly meaningful.

Because literal translations are not always the most accurate means of translating a concept, the result is that a back-translation may identify a wording that sounds unusual in the translated English. It may sound translated, and to a nonbilingual researcher, this would appear to be a problem when it may not be in reality. The flip side is that a poorly translated instrument may "pass" the back-translation test if it is a literal, word-for-word translation of the English. In short, back-translations defy the notion that cognitive equivalence trumps literal equivalence as a goal in translating and reviewing surveys. Thus, this practice can identify problems that do not actually exist and may not identify true problems with the translation.

The recommended alternative to the previous two methods is the *multicultural committee review* (Harkness, 2002; Schoua Glusberg, 2000). This process involves more than just two bilingual staff in the review. The committee members would ideally be selected to represent the various cultures present in the sample population. The review process can be handled in several different ways. The general idea is that each member has a chance to review the translation and the original English document individually, then the group is convened to discuss their individual concerns and decide as a group what the best solution is. Potential problems with this approach involve the possible domination of one committee member's opinion over another and the effect of the review process on the schedule and the budget. If these factors can be managed, the multicultural review can provide excellent insight into the varying ways of communicating effectively across Spanish-speaking cultures.

A common assumption of all of the translation and review methods discussed thus far is that all staff involved presumably would be well-educated bilinguals. That is, they speak, read, and write both languages fluently. As a result, their cognitive processing may have developed differently from the monolingual target population, who may be less facile with written language. Translators and reviewers must be aware of this discrepancy and must be conscientious about lowering the register of the translation so that respondents with less formal education also will be able to understand the questions.

For the 2000 survey's translation review, the aforementioned translation service provided a multicultural review of the existing translation of the survey instrument. Three translators participated, all native Spanish speakers experienced in survey translation and familiar with the committee approach to survey review. Two of the translators were males, one from Mexico, the other from Puerto Rico. The third was a female from Peru. The committee meetings were coordinated and refereed by a bilingual with extensive experience in social science research. Each translator was provided with a copy of the Spanish instrument in advance for review. The committee then met over several sessions in April and May 2000 and discussed any and all instances of possible awkwardness in the existing translation, as well as potential solutions for them.

The survey multicultural review, while not revealing any glaring errors in the existing translation, resulted in the discovery of numerous instances of awkward phrasing and word choice resulting from direct translations being used rather than cognitively equivalent translations. Through the use of focus groups, which are discussed in the next section, a large number of changes were recommended in the final report. As mentioned earlier, RTI also used in-house bilingual personnel who were familiar with the survey and the practical implications of making changes to it to review the findings and advise the researchers and study sponsors.

Testing the Translation

In addition to language proficiency issues, innumerable other cognitive, social, and environmental influences act on different individuals' understanding of the meaning and intent of survey questions. These issues are not fundamentally different from those encountered in the development of an English questionnaire. The solution to the problem for the Spanish translation is the same: Test it with "real" respondents. Ideally, testing would involve focus group reviews to reveal cultural and subcultural conceptual incongruities, and cognitive lab interviews in Spanish then would be used to obtain more detailed assessment and generation of precise wording choices. These are time-consuming processes, and practical considerations obviously will dictate which methods are used and for what purposes. However, these methods, when considered in the context of overall survey production, are inexpensive, and time should be allotted for them in the planning stages.

As with the selection of translators and reviewers, participants in the testing should be selected to represent, to the extent possible, the culture and dialect of the target population for the actual survey. Age, gender, and other characteristics also should be matched as closely as possible in order to provide the most accurate and informative review. Perhaps the most difficult aspect of the entire testing process is finding qualified, experienced focus group moderators and/or cognitive interviewers who also are fluent in both the target and the standard languages.

To test the findings of the committee review of the 2000 survey, four focus groups were conducted. The focus groups were held in the Chicago area, moderated by the expert bilingual person who earlier coordinated the translation review committee. One focus group was composed of adult Spanish-speaking drug rehabilitation participants; the other adult group was comprised predominantly of people of Mexican descent. There also were two groups of youths (12 to 17 years old), one predominantly Puerto Rican and the other predominantly Mexican.

The focus group participants were selected on the basis of personal characteristics that particularly qualified them to provide comments on specific sections of the survey or its supporting documentation. Moreover, the participants were selected to represent at least some of the diverse Spanish-speaking groups currently residing in the United States. The overall objective of the focus groups was to collect additional information regarding specific topics or questions that appeared to be problematic during the detailed review of the translation, which was completed earlier by the committee of translators.

After receiving consent from all participants, the sessions were recorded for the purposes of later clarification as necessary. The participants introduced themselves, using first names only, and provided basic demographic information about themselves, including age, country of origin,

and length of time living in the United States. The information regarding country of origin was critical in evaluating the responses provided by each participant. The moderator took this information into account as she considered the opinions of the individuals in the group, paying particular attention to the participants of Mexican descent because they represented the majority of the respondents completing the survey in Spanish.

In general, the findings from the four groups validated the modifications suggested by the committee, but there also were cases in which the original translated portions of the 2000 survey were determined to be the preferred option. The most striking observation of all the groups was that there was wide variability in the preferred way to say things in Spanish, and at times there seemed to be as many opinions as there were participants. However, the objective of the discussions was to attempt to determine the wording that is optimal for understanding by all individuals in the group. In this way, each group could come to a different conclusion regarding the best word choice based on their own cultural background, life experiences, or age.

The suggestions generated by interpretation of the focus group results and the earlier multicultural translation review then were reviewed again by RTI's panel of in-house translators, and a majority of the suggested revisions were adopted for the 2001 survey. In general, the review and focus groups were responsible for the production of a revised version of the Spanish instrument that used terms more likely to be recognized and understood throughout the diverse spectrum of Spanish-speaking respondents, and, when possible, used simpler language to convey equivalent meaning.

Table 6.1 displays weighted percentages of Spanish-language survey respondents, as rated by field interviewers (FIs), on how much difficulty they had with the interview. Data were combined for the 2 years before and the 2 years after the changes were implemented. There are obvious differences between the distributions from the two time periods, as the overall percentage of respondents having a lot of difficulty was cut in half (6.4 to 3.2 percent), while the percentage having no difficulty increased from 56.6 to 65.0 percent.

Table 6.1 Field Interviewer Ratings (Weighted Percentages) of Respondent Difficulty with Spanish Interviews in the 2 Years Before (*N* = 4,313) and After (*N* = 4,356) the Implementation of Changes

	Age Group							
	12 to 17 Years		18 to 25 Years		26 or Older		Total	
Difficulty Rating	1999 and 2000	2001 and 2002	1999 and 2000	2001 and 2002	1999 and 2000	2001 and 2002	1999 and 2000	2001 and 2002
None	67.5	79.8	63.5	72.9	54.8	62.5	56.6	65.0
A little	21.0	14.7	25.9	17.5	27.4	25.0	26.9	23.3
A fair amount	8.3	4.7	7.6	7.0	10.1	8.8	9.7	8.3
A lot	2.1	0.8	2.7	2.6	7.3	3.5	6.4	3.2
No response	1.1	0.0	0.3	0.0	0.4	0.2	0.4	0.2

Common Problems Encountered in Producing and Administering Spanish Instruments

The remaining text focuses on the most common problems that English-to-Spanish survey translations encounter. When possible, examples are drawn from the issues that were raised during the development and review of the Spanish CAI instrument for the 2000 survey. Solutions that were implemented by the review committee also are discussed, as are procedures that have been in place for some time as a result of experience in administering the survey in Spanish.

Cultural and Dialectical Issues within the Spanish Language

Contrary to popular myth among nonbilingual persons, there is no "standard" Spanish that can be applied to translations. Especially for national household studies in the United States, the diversity among Spanish-speaking respondents is significant. There are many instances in Spanish in which the words used in one dialect mean something quite different in other Spanish-speaking communities. Consider the following example:

> The word *cigarro* means "cigarette" to Mexicans. However, to Cubans, the word *cigarro* means "cigar."

This can be a serious problem for NSDUH in that it seeks to gather information regarding individual types of tobacco products. One solution to this problem is to use multiple word choices within the survey instrument, but this can be awkward and confusing when the question is complicated with alternative word choices. Another solution that translators frequently use is to describe the item or concept so that all Spanish-speaking respondents are more likely to understand the intent of the question. Because of the importance of clear communication in this introductory sentence, the solution adopted by the review committee incorporated a combination of these two approaches. The revised lead-in to the tobacco module explains,

> *Estas preguntas se tratan del uso de productos de tabaco. Esto incluye cigarrillos, tabaco de mascar, tabaco en polvo (rapé o "snuff"), cigarros (puros) y tabaco en pipa. Las primeras preguntas se tratan solamente de cigarrillos.*
> "These questions are about your use of tobacco products. This includes cigarettes, chewing tobacco, snuff, cigars and pipe tobacco. The first questions are about cigarettes only."

The term *cigarillos* was used to distinguish cigarettes from cigars–*cigarros*–using both terms within the same sentence and thus increasing the salience of the comparison. Parenthesized terms *(rapé, puros)* also were used to clarify word meanings in this introductory statement.

Cognitive Equivalence Versus Literal Equivalence

The typical target audience may not understand translations that are too literal. A bilingual translator is capable of cognitively processing both the Spanish words and the English sentence structure and semantics so that the intent of the question seems clear even if the wording is awkward. Although a literally translated question may be meaningful to a bilingual translator, that person's thought patterns are different from those of monolingual respondents, who are

unfamiliar with English grammar and meanings. The best illustration of this potential problem is to reflect upon the difficulties one encounters when trying to understand someone who is not a native speaker and does not speak a language fluently. Beyond the pronunciation difficulties, there typically will be sentence construction and word choices that do not sound natural even if the grammar may be technically correct. Consider the following example:

> The English version of the 2000 survey contains a question that asks, for a carton of cigarettes, "What was the price you paid?" Prior to review, this was literally translated as *¿Cuál fue el precio que usted pagó?* This question would be better understood by most Spanish speakers as *¿Cuánto pagó*? or "How much did you pay?" (which was the solution adopted by the review committee).

Sometimes an exact or direct translation of a concept can be awkward or may not be meaningful to all respondents in the target language. This can allow error into the data unnecessarily.

Educational Level and Respondent Literacy Issues

It is quite possible to find lower educational experience and reading skills among newer immigrant groups, particularly those who have left their native countries as a result of poverty or political repression. Sometimes, it is necessary to use multiple terms (word alternatives) in Spanish, or even English terminology, to ensure communication with all segments of the population. In some cases, a word may be correct, technically speaking, but may not be understood by respondents with lower reading skills. For example:

> The term *hijo de crianza* in Spanish is the correct term for "foster child." However, the term *hijo adoptivo temporal* is better understood by some respondents.

In this particular case, there is a cultural issue that complicates the matter. The concept of a "foster" child or parent does not exist in Latino culture, in which it is more likely that another family member would accept a child into an extended family. The best recommendation in situations of this sort is to include the English term as well as the descriptive term in the question. Most likely, a Spanish-speaking respondent who is involved in the foster care program in the United States would recognize the English term, and the Spanish description would help to prevent incorrect responses. The term that was agreed upon following the translation review was *hijo "foster" (de crianza).*

Gender Issues in Spanish

It is common knowledge that there are gender issues in Spanish that are nonexistent in English. Male respondents must be addressed differently from females. Accommodating both genders, that is, using tailored scripts, when possible, avoids the problem of awkwardness caused by including both masculine and feminine forms. It also ensures that the respondent is not insulted or confused by the question wording if both options, or only the masculine, are used.

Although paper-and-pencil questionnaires are limited with regard to wording (typically, questions are addressed to males with the feminine endings appearing in parentheses), the CAI mode facilitates tailored scripting. The program can take the gender into account and only display what is appropriate for the respondent. However, this means that the programmer must take additional time to create the logic to accomplish this, and the sound file must be recorded twice. Consider the following question, which is a common demographic question in surveys, "Are you currently, married, widowed, divorced, separated, or never married?"

> *Actualmente, ¿está usted casado(a), es viudo(a), divorciado(a) o separado(a), o nunca se casó?*

The computerized solution would involve constructing a gate question for the interviewer to enter the gender of the respondent. Then, the computer chooses and displays the question as written appropriately for persons of that gender.

Due to the extensive resources required to develop and store multiple versions of each question and answer in Spanish, NSDUH currently provides only the masculine wording for the item above in the audio files. It uses gender-neutral terms in other situations, in spite of their being grammatically incorrect. These infractions can be seen as roughly equivalent to using "they" when "he or she" would be correct. The review committee recommended that several of these items should be corrected. With the laptop upgrade that was rolled out for the 2004 NSDUH, there will likely be sufficient computing resources to maintain, where necessary, two Spanish audio files and the required skip logic to make the Spanish instrument grammatically correct in all cases.

Formal and Informal Verb Tenses in Spanish

In addition to gender issues, Spanish has a formal and informal subject and verb tense for the second-person singular. Based on the age of a respondent and the interviewer's familiarity with or relationship to the respondent, either the *Usted* or *Tu* subject and verb tense would be appropriate for any question with the subject "you."

Using the informal verb tense and article (*Tu*) could create tension or possibly even inhibit cooperation with adult respondents because it is a presumed intimacy that may not exist. In survey research settings, to do so with adult respondents is rarely appropriate. However, the informal tense *is* appropriate for youth respondents. This means that there would need to be two different questions (or fills) for Spanish to accommodate respondents of different ages. Again, this complicates the programming and takes up more memory on the laptop computer, but it can be handled through the programming of gate questions. The researcher must work with the translator to decide how this is best addressed given the number of youth interviews expected and the impact of not using the correct terms. If it is necessary to simplify the instrument by selecting one verb tense, it is better to err on the side of using the formal tense (*Usted*).

When to Translate and When to Keep the English

This is a key issue and one that is not always easy to resolve. The researcher needs to understand the impact of the decision, so it is incumbent on the translator to explain the options

available. In a previous example, it was recommended to include the English term for foster child along with the Spanish description and possibly the technically correct term. In another example from the survey instrument, questions related to alcohol consumption contain specific names of alcoholic beverages that are in both English and Spanish. Decisions must be made to determine what is appropriate. For example:

> In English, one would not translate the drink, *Tía María to "Aunt Mary*." Therefore, the Spanish version should not include a translation of the drink, "Bloody Mary" to *María Sangrienta*. Not only is it too graphic, it is not an alcoholic drink of any sort. English names are consequently retained for many of the drinks named in the list of alcoholic beverages shown to NSDUH respondents.

Is "Spanglish" Ever Appropriate?

Researchers are reluctant to use incorrect grammar and wording, but sometimes it is necessary in order to be understood by the respondent population. In extreme cases, a false cognate is the most common way that a concept is communicated in Spanish and may be the best way to express it in a survey. Consider this example from NSDUH, wherein respondents are asked several questions regarding drug treatment:

> The term *consejería* is not a real word, but it is understood to mean "counseling" among Spanish speakers in the United States.

The term *consejos* is a real word in Spanish, which means "advice." Adding the ending *-ería* to a noun makes it a place where something is made, done, or takes place. The term *consejería* seems like it should be a place where one gets counseling, but it is simply not a real word. Spanish-speaking populations in the United States usually understand this to be counseling, and lacking a more appropriate alternative, the term is currently used in the NSDUH instrument.

Questions with Programmed "Fills" That Do Not Always Work the Same Way in Spanish and English

Because the sentence structure of Spanish is not identical to English, it is not always possible to have the identical programming fills (i.e., information that is added to a survey question, such as the date or respondent characteristic). Dates, for example, may appear in different parts of the sentence, and verb tenses do not always change consistently in both languages. Sometimes, it is necessary to program "empty" fills in one language or another because the sentence structure is not the same. The main problem with this is that the programmers do not know where to break the text for the fill if they are not bilingual in Spanish. The example below is a sentence with fills that are different in English and Spanish.

> Now think about the past 12 months, from [July] [28] [1999] through today.
>
> *Ahora piense en los últimos 12 meses, desde [el 28] [de] [Julio] [1999] hasta hoy e incluyendo el día de hoy.*

The solution to this problem that the 1999 survey used was to provide completed programming code to the translator and teach him or her to enter the Spanish text just as the English appeared. The translator constructed the stem questions as well as the "fills" so that the translation sounded natural. Obviously, this was much quicker and easier than teaching the programmer to speak Spanish.

Cultural Differences in the Reporting of Dates and Age

There are subtle differences in language that could easily be overlooked if the translator and researcher are not careful. In determining the manner in which a date is to be reported, it is necessary to be very specific regarding which part of the date is recorded first. The following example illustrates the distinction in reporting dates for English-speaking U.S. respondents and many Spanish-speaking (or European) respondents.

> 5-10-00 is May 10, 2000 in English for respondents in the United States. However, it is October 5 to a native Spanish speaker.

Another interesting cultural difference in reporting involves the respondent's age. In many Spanish-speaking cultures, it is common to report age on the upcoming, as opposed to most recent, birthday. Although the age question may be translated faithfully, the incorrect age may be reported and go unnoticed by interviewer and respondent alike. Here is an example:

> English question: "How old are you?"
> Spanish question: *¿Cuántos años tiene?* (Same question)
>
> English response: "I'm 35 years old."
> Spanish response: *Voy a cumplir 36 años.* (I will be 36 years old.)

Without an awareness of this issue, the result could be inaccurate data being collected. The solution adopted by NSDUH for this issue is to ask for the year of birth and then confirm the age. The first question asked is "What is your date of birth?" (which is accompanied by a note to the interviewer to enter MM-DD-YYYY). The next question that appears is "I have entered your date of birth as [month - date - year]. Is this correct?" Then the computer calculates the age and reconfirms, "That would make you __ years old. Is this correct?" In this manner, the interviewer almost certainly records these data as intended.

Selecting a "Voice" for ACASI Recording

Once all of the precise wording choices have been finalized, the audio recording for ACASI sections can begin. At a minimum, the goal is to find a voice that is understood by all Spanish-speaking populations. Other factors that could be considered involve potential biases associated with perceived age, gender, or race/ethnicity of the voice. However, there have not been sufficient studies done on these factors to date. For this reason, the minimum requirements of being understood by all Spanish-speaking populations, and matching the demographics of the corresponding English voice, should be the primary focus in selecting a Spanish ACASI voice.

Common sense dictates that a voice with an American accent is more likely to be difficult for a non-English-speaking person to understand and could potentially bias the responses for some questions. Cuban or Puerto Rican dialects may be difficult for some respondents to understand because these dialects are unique and tend to be spoken very quickly. It would be a mistake to assume that a voice from Spain would be best for a survey in the United States because this accent also is unique, and there would be very few respondents who originate from Spain among U.S. Spanish speakers. Although a Mexican accent is likely to be more common among the survey participants, it too is a distinct accent that could be a challenge for other survey populations. Generally, a Colombian accent is well accepted by most other cultures and therefore would be a good candidate for an ACASI voice. The most important criterion, however, is that the recording be read as clearly as possible, regardless of the accent.

Beyond the issues related to accents, a "professional" voice, like a singer, actor, public speaker, or radio/TV personality, will improve the overall quality of the recording and will likely improve the speed with which the process is completed. Well-trained interviewers also are easy to work with and may get the recording right in fewer attempts, and they also may be more affordable than other professional voices. Respondents are likely to prefer a voice that is easy to listen to as well as understand. Very unusual voices could be distracting or even irritating to respondents. An additional consideration is the likely availability of the speaker for producing future updates of the survey as questions are revised over time and new modules are added.

The voice used for the ACASI sections of the Spanish NSDUH was selected on the basis of gender and age matching to the corresponding English voice. Based on the results of cognitive laboratory testing, the English voice chosen for NSDUH is female, with somewhat lower pitch than the other female voices tested, and with comparatively little volume dynamics. The Spanish voice was matched to the English voice to the extent practical and is that of a female native Colombian who is herself a translator and public speaker.

When to Translate Interviewer Instructions

There are several schools of thought regarding this issue. Some researchers and translators want everything translated, and some do not want any interviewer instructions translated. The best answer actually depends on the situation. There are times when both are appropriate. If interviewer instructions are not translated, interviewers may be confused by having to switch back and forth between languages. If the instructional text provides specifications for the respondent regarding the intent of a question, then not having this text translated means the researcher is relying on the interviewer to translate on the fly. This can be difficult for the interviewer and may introduce bias.

Sometimes, translating interviewer instructions is difficult because there may be survey jargon that does not translate into anything meaningful, which can cause more confusion. For instructions that contain survey jargon, like "Skip to Q3," it may make more sense to leave the instructions in English and not force the interviewer to have to decipher back-translations to figure out how the survey jargon was translated. Occasionally, usability testing may be called for in order to resolve differences of opinion on these types of questions.

Version Control Problems

Despite the best efforts of researchers, it is common to have tight schedules and last minute revisions to the English instrument. In the best case, this merely creates additional translation work if the translator has already completed the work before the English version changed. The worst case involves revisions to the English version that are not made in the Spanish prior to fielding. Version control can be a challenge to keep up with when iterative changes are made to the English, especially when ACASI use requires re-recording of individual items. The best solution is to maintain a timeline that always takes into account the additional effort required to translate, test, and produce the Spanish version.

A related issue is documentation. When changes are made to the translated instrument, it is imperative that the changes are recorded in such a way that they can be easily tracked. Reasons for the changes should appear as well to prevent ill-advised returns to previous wordings that proved unsuccessful in the past. Most importantly, the documentation should take place simultaneously with the changes (Harkness, 2002). Attempting to reconstruct decision processes at a later date is often fruitless.

Selecting and Certifying Bilingual Interviewers

Once all of the instrumentation issues have been settled, it may be difficult to find interviewers who are qualified to do their work in Spanish. While possessing all of the skills of nonbilingual interviewers, these individuals also must be able to communicate fluently in English as well as Spanish because a large proportion of their public contact will likely be with English speakers and because they are most often trained and supervised in English. Although English skills are readily recognizable by nonbilingual staff, a formalized certification process for bilinguals should involve reading comprehension, listening skills, and a test for conversational fluency in Spanish. Interviewers who cannot demonstrate proficiency in all three areas should not be allowed to conduct interviews in Spanish because it is up to these people to ensure that the meanings and intentions of the researchers are accurately conveyed to potential and actual respondents. The assessment should be designed in such a way that it would be an adequate measurement of the language skills for any research effort.

The method used to certify bilingual interviewers for NSDUH has been in place since 1994. RTI's Spanish Language Skills Assessment is administered to all bilingual FI candidates before they are hired to conduct interviews in Spanish. Reading comprehension, listening skills, and conversational fluency also are assessed by in-house translation specialists during new-to-project training. The record of the language skills assessment is maintained in a common file so that interviewers are not required to complete such an assessment for each survey for which they conduct interviews. The bilingual certification process has increased in importance with the growth in sample size of the survey in recent years.

Conclusions

This chapter has described recent efforts to update the Spanish translation of the NSDUH instrument and illustrated some of the dangers of inaccurate translations. In doing so, it presents a possible protocol that could help to solve some of the most common problems that researchers

and translators face. In summary, the multicultural translation review undertaken in 2000 examined a translation that was already considered adequate, and it fine-tuned many of the questions to make the instrument more understandable and less burdensome, both to respondents and interviewers. The revised questionnaire was introduced in the 2001 survey's data collection. The process resulted in the elimination of instances where direct, literal translations interfered with respondent understanding and replaced them with questions that were cognitively equivalent to the English versions and, therefore, were more readily accepted by the majority of Spanish speakers. Although the data presented earlier (***Table 6.1***) indicate that respondent understanding was probably improved after the results of the review were implemented, it can only be assumed that these changes led to the capture of more accurate data because "gold standard" comparisons are costly. Comparisons with data obtained from the English version, such as those using Item Response Theory-based analyses (e.g., Ellis & Mead, 2000), or factor analytic approaches (e.g., Rio, Quay, & Santisteban, 1989) might provide more insight into the probable success of the endeavor.

The demographics of the U.S. population are changing to include increasing numbers of non-English-speaking individuals who potentially may be selected to participate in a survey. For a national survey *not* to provide a means to include these respondents would be to underrepresent a portion of the population that is already likely to be somewhat excluded from public health, education, and other social welfare programs. Although it may not be possible to provide translations for every non-English speaker, effort should at least be directed toward the development of effective Spanish translations because of the relative size of this group. The intent then is to obtain the same high-quality data from English- and Spanish-speaking sample members. However, the conclusions drawn from field tests, focus groups, cognitive lab interviews, and other methodological studies that test the logistics or soundness of a survey in English do not transfer automatically to the Spanish-translated version. Therefore, the translation and review processes, wherein cross-cultural differences and cognitive-linguistic incongruities are ironed out, should be allotted a sufficient amount of time and resources. Indeed, if the researcher is interested in obtaining quality data that are at all comparable with those obtained from English speakers, these processes should receive at least as much time and attention as was devoted to the development of the original English version.

References

Behling, O., & Law, K. S. (2000). *Translating questionnaires and other research instruments: Problems and solutions* (Quantitative Applications in the Social Sciences Series No. 07-133). Thousand Oaks, CA: Sage Publications, Inc.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1*, 185-216.

Brislin, R. W. (1980). Translation and content of analysis of oral and written material. In H. C. Triandis & J. W. Berry (Eds.), *Handbook of cross-cultural psychology: Volume 2. Methodology* (pp. 389-344). Boston, MA: Allyn and Bacon.

Casagrande, J. (1954). The ends of translation. *International Journal of American Linguistics, 20*, 335-340.

Ellis, B. B., & Mead, A. D. (2000). Assessment of the measurement of equivalence of a Spanish translation of the 16PF questionnaire. *Educational and Psychological Measurement, 60*, 787-807.

Harkness, J. A. (2002). Questionnaire translation. In J. Harkness, F. J. R. Van de Vijver, & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35-56). Hoboken, NJ: Wiley.

Rio, A., Quay, H., & Santisteban, D. (1989). Factor-analytic study of a Spanish translation of the revised behavior problem checklist. *Journal of Clinical Child Psychology, 18*, 343-350.

Schoua Glusberg, A. (2000). *Translating research instruments: Committee approach and focus groups*. Chicago, IL: Research Support Services.

Shin, H. B., & Bruno, R. (2003, October). *Language use and English-speaking ability: 2000* (Report No. C2KBR-29, Census 2000 Brief). Washington, DC: U.S. Bureau of the Census. [Available as a PDF at http://www.census.gov/prod/2003pubs/c2kbr-29.pdf]

7. Analyzing Audit Trails in NSDUH

Michael A. Penne and Jeanne Snodgrass
RTI International

Peggy Barker
Substance Abuse and Mental Health Services Administration

Introduction

In 1999, the data collection method for the National Survey on Drug Use and Health (NSDUH) was changed from paper-and-pencil interviewing (PAPI) to computer-assisted interviewing (CAI).[1] The interview sections on substance use and other sensitive topics were changed from self-administered PAPI to audio computer-assisted self-interviewing (ACASI). These changes were prompted by research that showed that CAI questionnaires reduced input errors; research also showed that use of ACASI increased comprehension for less literate respondents, and, by increasing privacy, resulted in more honest reporting of illicit drug use and other sensitive behaviors (Lessler, Casper, Penne, & Barker, 2000). Early research on the survey demonstrated the value of using keystroke files (predecessor to audit trail files) to evaluate the ease with which respondents navigated through the ACASI portions of the questionnaire (Caspar & Couper, 1997). In this chapter, the earlier work is briefly described, possible methods for streamlining the data-processing portion are discussed, and audit trails in the 2002 survey are used to investigate three aspects of data quality: question timing, respondent breakoffs, and respondent "backing up" to change prior responses.

In preparing for the 1999 conversion to CAI, a field test was conducted in the fall of 1996 with 435 respondents to compare two CAI versions of the 1996 PAPI instrument. As part of this experiment, an MS-DOS Blaise program was modified to capture each keystroke made, including function keys, during the course of the interview by both the respondent and the field interviewer (FI). Additionally, the program captured timing data for each screen encountered and an indicator of whether the audio or screen component was turned on or off. Analysis of these keystroke files centered on the respondents' interaction with the ACASI program. Keystroke file analysis from both CAI versions indicated that roughly 60 percent of all respondents utilized a function key at least once with very little difficulty and overall that respondents would not experience any difficulty while interacting with the ACASI portion of the instrument (Office of Applied Studies [OAS], 2001).

Simultaneous with the conversion of the questionnaire to CAI for the 1999 survey, the Blaise programming language changed from an MS-DOS-based version to a Windows-based version,

[1] Prior to 2002, the survey was called the National Household Survey on Drug Abuse (NHSDA). In this report, it is referred to as NSDUH, regardless of the survey year.

and the keystroke file application was no longer compatible. This new version of Blaise did, however, contain code that would capture data similar to the previous keystroke file application. These new files were called "audit trails." It should be noted that unlike the keystroke file application, which recorded every keystroke ever pressed, including deletions, the audit trail file application only records the final response entered before the respondent moves to another screen. To illustrate this difference, consider the following example: Under the keystroke application, if a respondent initially entered a value of "1" ("Yes"), then backed up, deleted the "1" value and entered "2" ("No") instead, the keystroke file would record a value of "1 [backspace] 2" for that particular screen. The audit trail file on the other hand would only record a value of "2."

Interest in capturing this type of data was renewed after the first nationwide administration of the CAI instrument in 1999. Issues surrounding changes in data collection mode prompted renewed interest in audit trail file analysis, and the decision was made that some portion of audit trail files should be retained for analyses. Because a sample size of approximately 25,000 audit trail files would be more than adequate to answer any proposed questions, one in every three transmitted audit trail files was retained starting with the 2000 survey year.[2]

Analysis of the 2000 CAI instrument audit trail files was done to validate the findings from the 1996 keystroke file analysis. Sample sizes for this analysis incorporated a little over 12,500 audit trail files (one third of the data from two calendar quarters). The results validated the 1996 findings that respondents were having little difficulty completing the ACASI portion of the questionnaire (Caspar, 2000).

Through the 2000 survey year analysis, unexpected problems were encountered with the retention of one in every third transmission methodology. Most notably, when a breakoff occurred within an interview but was completed at a later date, there were instances where the last portion of an interview was retained, but not the initial portion, and vice versa.[3] As a result, the level of partially completed interviews for audit trail analysis was high, which made it difficult or sometimes impossible to use the data. This problem was large enough, and the interest for further data quality analysis was strong enough, to warrant retaining all transmitted audit trail files. Conclusions about the difficulties associated with retaining one in every three transmitted files were not made until well into the 2001 survey year. Because it was not feasible to modify the data collection process to retain all transmissions at that time, the change was implemented at the beginning of the 2002 NSDUH year.

In 2002, attention turned toward an investigation of audit trail analysis to use for possible data quality measures. To assist in preparing for this analysis, the 2001 audit trail data were used to plan, develop, and test both the analysis methods and the programs. Then the first 6 months of audit trail data from the 2002 NSDUH were used for testing (Penne, Snodgrass, & Barker, 2002).

[2] A single transmitted file does not imply either a single interview only or an interview in its entirety. Field interviewers (FIs) are required to transmit any completed work nightly for every day that they work. This implies that a single transmission may contain more than a single interview.

[3] Because FIs were required to transmit any completed work nightly, if an interview experienced a breakoff, the first portion of the interview would reside in one transmission file while, assuming that the interview was eventually completed at a later date, the remaining portion of the interview would reside in an entirely different transmission file.

Current Issues and Methods

All current analysis methods and results presented in this chapter are from the entire 2002 NSDUH. This analysis includes all completed and partially completed interviews, regardless of final response or usable status. An interview is classified as "usable" if the respondent provided data on lifetime use of cigarettes and at least nine other substances (OAS, 2002). It was felt that because many of the analysis issues are highly correlated with respondents' behavior, all data should be used. Prior to analyzing the 2002 NSDUH audit trail data, each aspect for examination was outlined. These aspects include question timing and changes in response, along with questionnaire breakoffs.

Timing

As was the focus of the timing analysis in the earlier studies of keystroke/audit trail data, the initial purpose for this analysis was to monitor respondent difficulties in utilizing ACASI. Though not presented in this chapter, the results were the same as before: Respondents were not having any more than expected difficulties in navigating the questionnaire. This reassurance of respondent behavior permitted redirecting the focus to FI performance and what possible ramifications it might have on respondent behavior.

Six measures within the instrument were chosen to examine FI behavior with regard to timing (see ***Table 7.1***).

Table 7.1 Measures Used in Audit Trail Analysis

Modules of Interest	Screen Name	Description
Introduction to CAI	INTROCAI	The interviewer begins rapport with the respondent and reads aloud the introduction to the study, if applicable.
Calendar Setup	CALENDAR	The FI sets up the reference calendar for the respondent to use throughout the interview.
ACASI Setup	INTROACASI	The FI explains the ACASI setup to the respondent, adjusts the headphones, and starts the respondent on the tutorial.
End of ACASI	ENDAUDIO	The respondent finishes the ACASI portion and turns the laptop over to the FI, who enters a three-letter code to continue.
Verification Form Completion	TOALLR3I	The FI interacts with the respondent to complete the verification form.
Ending Interview with Respondent	INCENT01	The FI ends the interview with the respondent and gives the respondent a cash incentive for his or her participation.

ACASI = audio computer-assisted self-interviewing; CAI = computer-assisted interviewing; FI = field interviewer.

These single question measures were present in every interview (unless there was an interview breakoff), as opposed to all other questions in the CAI that are asked dependent upon skip logic. It was felt that each of these measures would provide insight into how long FIs are spending on these important aspects of the questionnaire. Audit trail files capture both the date and time (hours, minutes, and seconds) that an individual (FI or respondent) enters and exits all encountered screens. The difference between these two moments is calculated and converted to time in minutes (the desired unit of measure). Based on the precision of this recording device, the smallest feasible time calculated is 1/60 = 0.0167 minute or 1 second. For these analyses, the initial step entailed calculating a gold standard (GS) timing for each of the six measures mentioned above. The GS measures are considered to be the appropriate amount of time a respondent spends on a question to fully comprehend the information and respond to that

particular screen. The GS measures were developed at RTI with three NSDUH staff members administering the NSDUH CAI instrument to three non-NSDUH staff members who ranged in age from 27 to 32. The interview files then were sent to a Blaise programmer, who calculated the timing for each of the three interviews. In all six measures, the highest timing across the three interviews was taken as the GS. Actual timing for the six variables from all completed interviews then was compared with the GS measures to identify FIs falling below this amount of time. This provided a sense of the magnitude of possible shortcutting.

Breakoffs

In the breakoff analysis, the goal was to explore (1) the magnitude of breakoffs overall, (2) whether any particular section of the interview had a large number of breakoffs, possibly indicating that respondents were having trouble or were offended by the questions, and (3) whether any particular interviewers accounted for a larger than expected percentage of the overall number of breakoffs. All instances were examined where a breakoff occurred within the interview, including the overall number of occurrences (total interviews with at least one breakoff), the number of breakoffs associated with individual FIs, and the location within the interview of the breakoffs. With regard to location, breakoffs were classified into three main categories: ACASI, FI-administered sections, and instances where the CAI system itself crashed. The ACASI sections were categorized further as tutorial, main core drugs (including alcohol and cigarettes but excluding nonmedical use of psychotherapeutic drugs), psychotherapeutic drugs, and noncore sections. FI-administered sections were categorized as those with some respondent interaction (instances where the FI is actually administering the questionnaire to the respondent) versus no respondent interaction (e.g., confirmation of FI identification number, FI debriefing questions).

Changes in Responses

When the survey was conducted in PAPI, respondents were asked every follow-up question in most of the core drug modules, even if they indicated they had never used a particular drug. Data editing procedures used answers from all follow-up questions to classify a respondent as a drug user or nondrug user.[4] This, in effect, gave respondents multiple opportunities to report that they had used a drug, even if their initial response was a denial. With the switch to CAI, routing logic was implemented that skipped the respondent out of a drug module if he or she indicated at the beginning of the module that he or she never used a particular drug. With this implementation, the PAPI editing procedure was no longer possible. However, the audit trail data permitted the identification of instances where a respondent returns ("backs up") to a previously answered question and changes his or her answer. The goal was to know how much drug use estimates could increase if it were assumed that if a respondent ever indicated "Yes" to using a drug, he or she was in fact a drug user. The first step was to determine if in fact there were a sufficient number of instances where this occurred to have an impact on drug use estimates. Because true reasons for a respondent changing his or her answers are speculative at best, a worst case scenario approach was taken to the analysis. In other words, it was presumed that all

[4] Because there were no skip patterns within most sections of the PAPI, a respondent was directed through all follow-up questions about their use of a particular drug, regardless of the respondent's actual use. Any indication of use of a particular drug classified the respondent as at least a lifetime user of that substance.

changes in answers were reflective of the respondent's desire not to disclose his or her true drug usage. With this in mind, the audit trails were examined to find instances where a respondent at any time indicated a positive response to using a particular drug, but then changed this response to "No." As noted earlier, if a respondent initially answers "Yes" but deletes that answer and responds with a "No" in the same screen, the audit trail only captures the final "No" responses. Hence, these types of changes cannot be accounted for in these analyses.

There were two aspects of interest for this analysis. First, all individual lifetime use gate questions[5] were analyzed to calculate the overall effect on lifetime drug use prevalence. Second, the effects that these changes might have on past year and past month use of a drug were examined. In this analysis, the goal was to determine the most recent period ever that a respondent indicated he or she had last used a drug regardless of his or her final response. For example, if a respondent initially reported past 30-day use of a drug, but later changed the response to past year, this analysis would finalize that respondent as a past 30-day user.

Data Management

As noted earlier, starting with the 2002 survey year, all transmitted audit trail files were retained. Unfortunately, this did not alleviate all of the data management difficulties. Specifically, the potential for the same FIs to replicate the same unique identification (ID) number for different individual interviews, or for two separate FIs to use the same ID, still existed. Additionally, because the audit trail analysis is currently separate from the questionnaire data analysis, corrections of incorrect ID numbers and subsequent linking of follow-up portions of an interview from an earlier interview breakoff were not applied to their respective audit trail files. Ultimately, this resulted in approximately 6.2 percent (~4,457 records) of the 2002 survey year records having at least one duplicate ID represented on the analysis file. For processing expedience, and because a sufficient sample size was still retained to produce reliable estimates, it was decided to remove these records with duplicate IDs from the timing, lifetime, and recency analyses (n = 63,811). However, all records were considered within the breakoff analysis (n = 68,268).

Results

The timing data in ***Table 7.2*** shows that when measured against a GS time, FIs are spending approximately the correct amount of time with the very beginning of the interview at the introduction to the CAI instrument screen. However, once past this point, they spend less time than the GS on several important aspects of the questionnaire, such as setting up the calendar, setting up the ACASI tutorial, completing the verification form, and ending the interview with the respondent. Conversely, they are taking longer than the GS in ending the ACASI portion of the interview. There are at least two possible reasons for these results. The first might be that the FIs are not performing their jobs correctly. The shorter times might indicate that FIs are reading

[5] Hallucinogens, inhalants, and prescription-type psychotherapeutic drugs (pills) each contain multiple gate questions. Each question focuses on particular substances classified as a hallucinogen, inhalant, or pill. For instance, hallucinogens ask questions on LSD, PCP, peyote, angel dust, mescaline, psilocybin, "Ecstasy," and any other hallucinogen not already asked about.

Table 7.2 Unweighted Audit Trail Timing Analysis: Respondents Aged 12 Years Old or Older, 2002 NSDUH

Modules of Interest	Gold Standard Time (Minutes)	Respondents 12 or Older (*n* = 63,811)		Respondents 12 to 17 Years Old (*n* = 22,221)		Respondents 18 to 49 Years Old (*n* = 36,645)		Respondents 50 or Older (*n* = 4,945)	
		Median	Percent Below Gold Standard	Median	Percent Below Gold Standard	Median	Percent Below Gold Standard	Median	Percent Below Gold Standard
Introduction to CAI	0.10	0.10	49.81	0.08	53.32	0.10	48.47	0.17	43.88
Calendar Setup	1.55	1.22	67.93	1.23	67.34	1.20	68.89	1.27	63.42
ACASI Setup	2.23	1.48	74.62	1.51	74.21	1.45	75.95	1.70	66.63
End of ACASI	0.20	0.48	3.97	0.50	2.78	0.47	4.30	0.48	6.86
Verification Form Completion	0.42	0.08	75.87	0.07	75.61	0.10	76.26	0.10	74.07
Ending Interview with Respondent	0.52	0.05	84.98	0.05	84.42	0.05	85.44	0.05	84.11

ACASI = audio computer-assisted self-interviewing; CAI = computer-assisted interviewing.

through the questions too quickly for the respondent to understand or to get a question's full meaning.

The second possible reason could lie with the method of the GS calculation itself. Because GS timing was calculated using simulated interviews at RTI, it did not capture the situations that affect timing in many real field situations. The GS calculation, then, could be seen as a little higher than what is realistic. Another limitation to the GS calculation is that it was not calculated by age. The ages of the mock respondents (27, 29, and 32 years) were not representative of the entire NSDUH population and may not serve as a realistic benchmark measure across respondent age groups. Means for determining the validity of these GS measures would entail either modifying the measuring criteria for each aspect or conducting field tests, with an adequate sample and appropriate age range. On another note, the longer time taken to end the ACASI portion of the instrument could simply be accounted for by rapport between the interviewer and the respondent. At this point in the interview, the computer is handed back to the interviewer, and the interviewer may stay on this screen (ENDAUDIO) while talking to the respondent, or answering any of the respondent's questions about ACASI.

Though the reasons for these timing results are somewhat speculative and GS times require additional validation, a trend across the analysis age groups was noted. As the age group of respondent goes up, there was a decrease in the percentage of FIs below the GS time. This indicates that FIs are taking more time with older respondents, who tend to be less comfortable with the CAI.

To further assist in drawing proper conclusions about the interviewers' time to complete, graphical distributions, including a mark for the calculated GS time, were produced of the audit trail timing data for the 12 or older population.[6] ***Figures 7.1*** to ***7.6*** were initially created depicting the actual calculated minute value themselves. This resulted in graphs that were highly skewed to the right[7] and provided no discernible information about the distribution. To assist in resolving this aspect, log transformations of the minute values were calculated and plotted instead. Graph tick marks represent integer values of the log-scale, and labels were reset to reflect the time in minutes[8] for ease of interpretation. The bars or bins between each tick mark represent 10 equal intervals on the log-scale (e.g., each bar between 0 and 1 [or 1 minute and 2.7 minutes, respectively] represent 0.0 – 0.09, 0.10 – 0.19, ……, 0.90 – 0.99). It is important to note that empty bins at the left end or low time to complete side of the graph are expected due to the precision of the measure (i.e., the smallest calculated times possible are 1 second, 2 seconds, 3 seconds, and so forth, which translate into respective log transformation values of -4.09, -3.40, and -3.00 and hence bars between these values will not be observed).

[6] Initially, graphs were produced for each of the six timing measures for each age group (including an overall graph). Upon careful inspection, it was noticed that only negligible differences in the shapes of the graphs among the different age groups. The only noticeable differences occurred in the magnitude of the distribution, which was expected as a result of the varying sample sizes in each age group. Hence, only results for the 12 or older population are presented.

[7] To illustrate, the range for ***Figure 7.1*** was 98.23 minutes with a minimum calculated time of 0.0167 minute and a maximum time of 98.25 minutes.

[8] For example, the log transformation -4 is e^{-4} = 0.0183 minute or 1.1 second and e^{o} = 1 minute.

Figure 7.1 Unweighted Distribution of Audit Trail Timing Data: Introduction to CAI (*n* = 63,811)

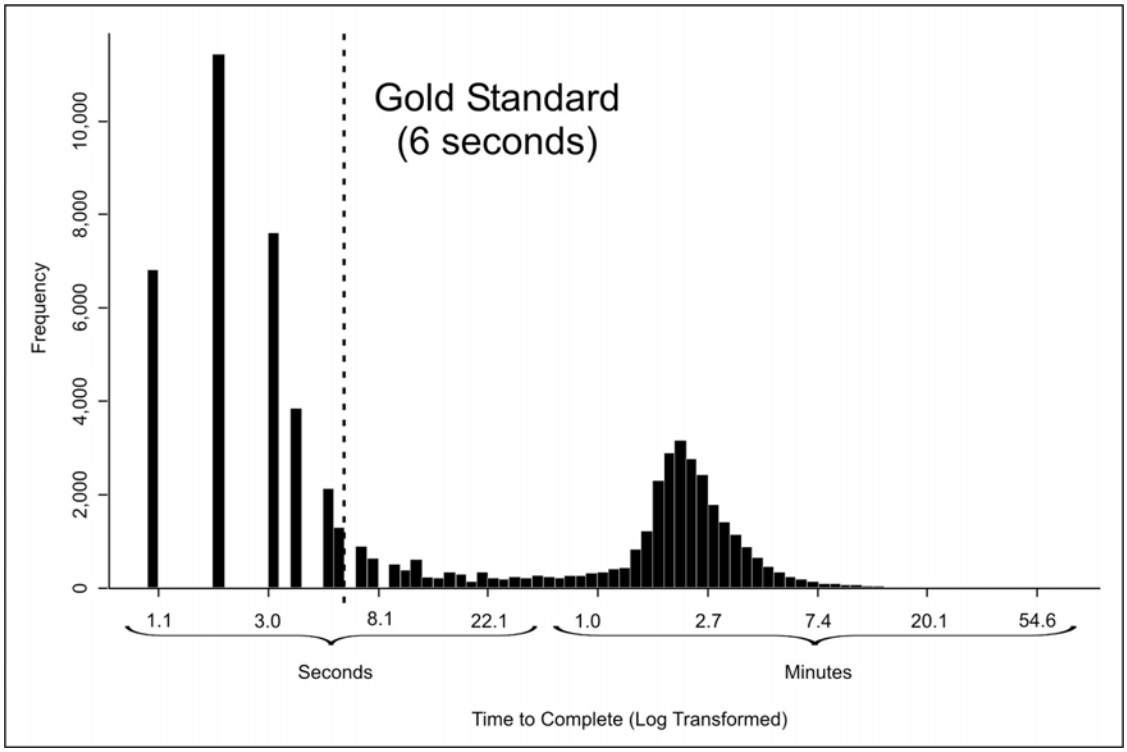


Figure 7.2 Unweighted Distribution of Audit Trail Timing Data: Calendar Setup (*n* = 63,811)

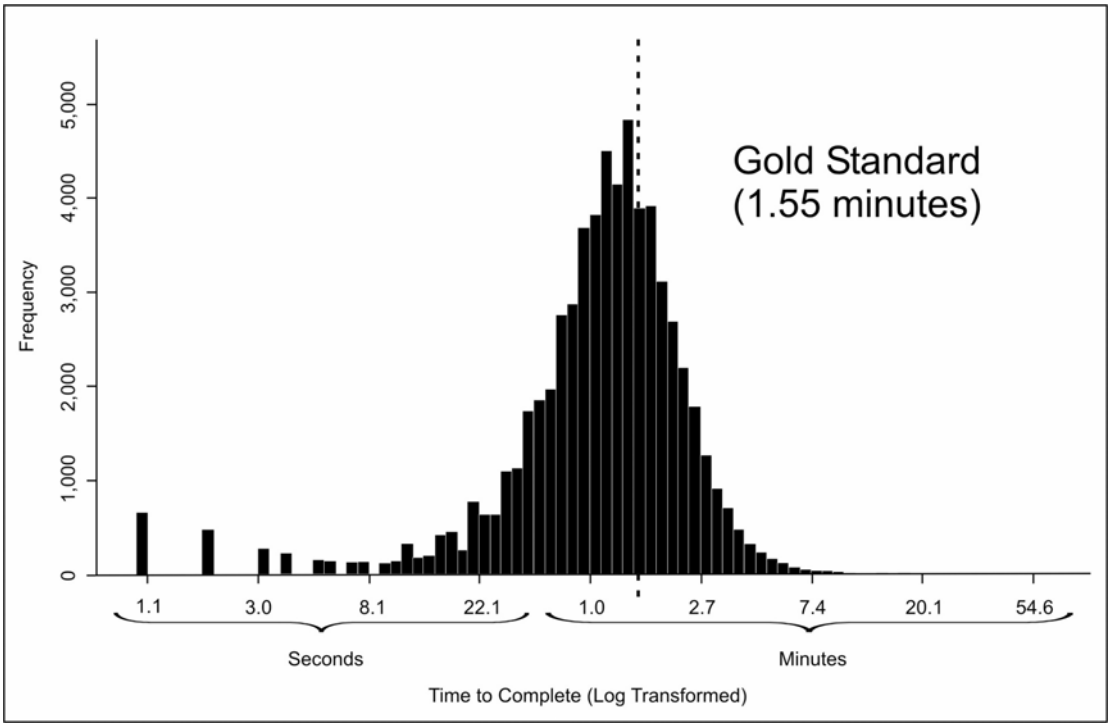


Figure 7.3 Unweighted Distribution of Audit Trail Timing Data: ACASI Setup (*n* = 63,811)

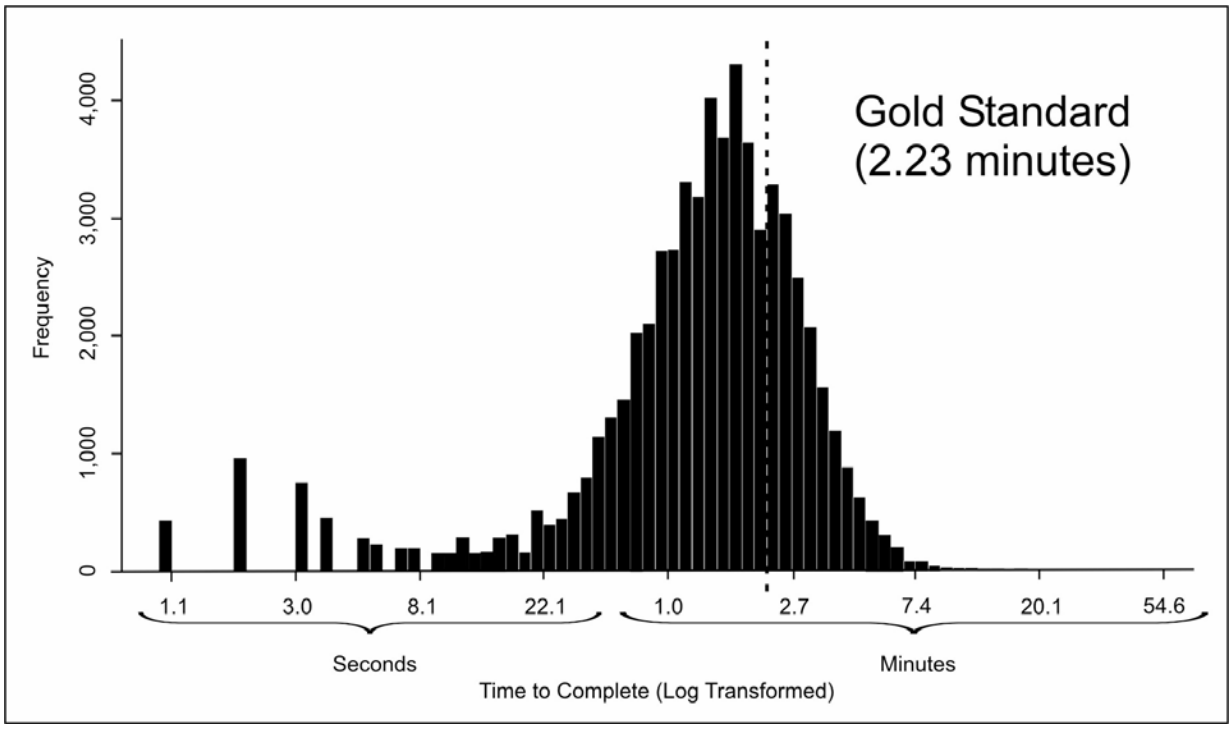


Figure 7.4 Unweighted Distribution of Audit Trail Timing Data: End of ACASI (*n* = 63,811)

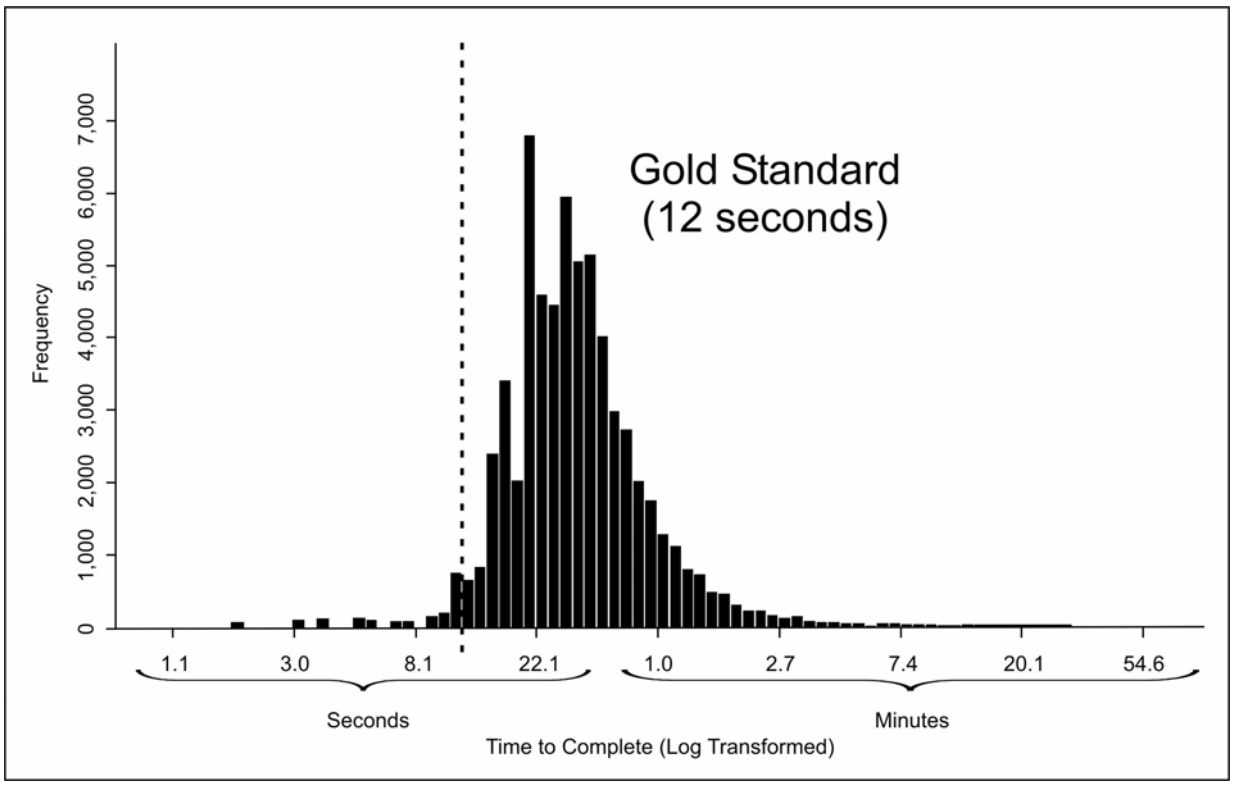


Figure 7.5 Unweighted Distribution of Audit Trail Timing Data: Verification Form Completion (*n* = 63,811)

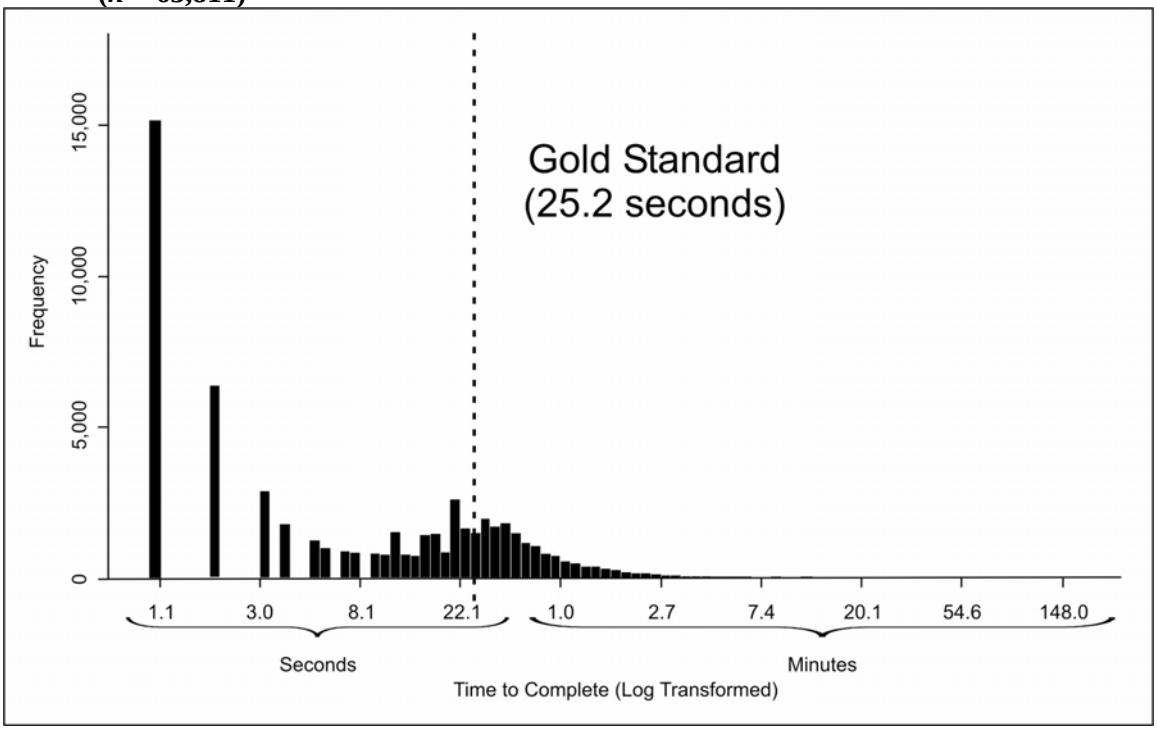
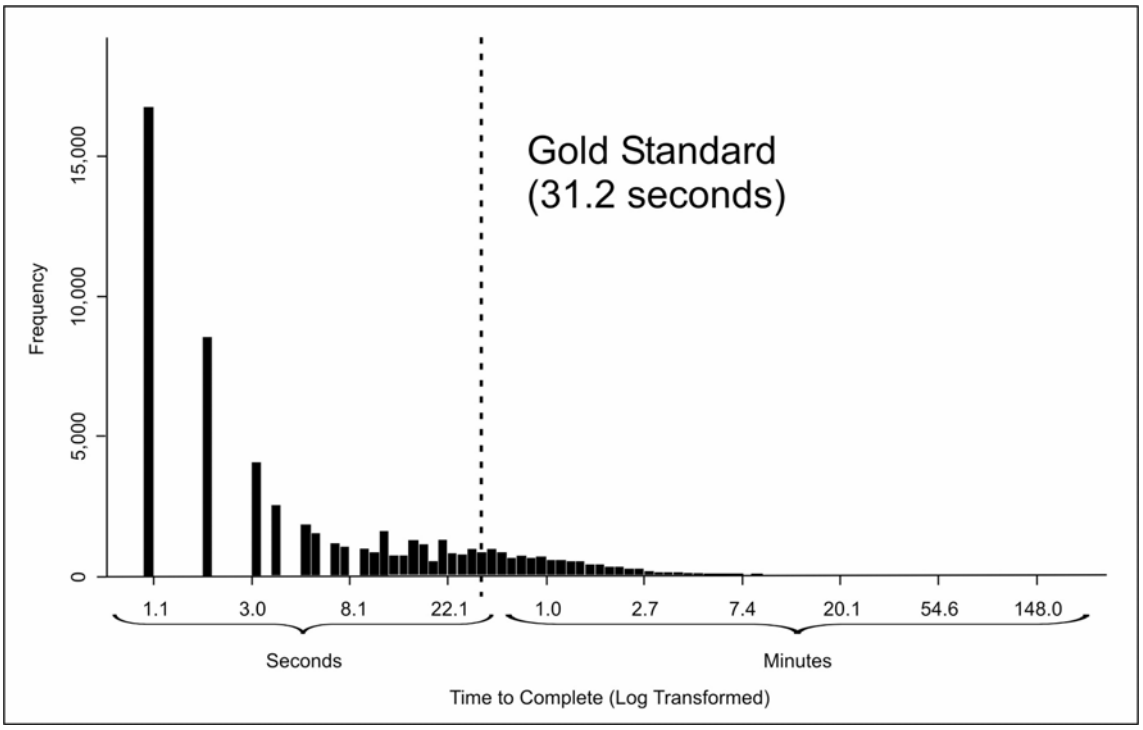


Figure 7.6 Unweighted Distribution of Audit Trail Timing Data: Ending the Interview with the Respondent (*n* = 63,811)



These figures show details that ***Table 7.2*** does not. For instance, in ***Table 7.2***, it appears that the interviewers were equally distributed around the GS for the introduction to the CAI, but from ***Figure 7.1*** it is seen the distribution is in fact bimodal. Further, roughly one third of the sample is taking less than 2 seconds to complete this screen (one third of the GS time). These same patterns are seen in later screens, also (verification form completion in ***Figure 7.5*** and ending the interview with the respondent in ***Figure 7.6***). These results indicate that shortcutting by the interviewer is a possibility. The data in ***Table 7.2*** and ***Figures 7.1*** to ***7.6*** are being investigated further by NSDUH data quality staff, along with many other indicators of data quality, for flagging interviewers whose work may warrant a closer look.

Tables 7.3 and ***7.4*** illustrate that interview breakoffs within the ACASI are minimal. It does not appear that there are specific areas in the questionnaire where respondents are breaking off, and the majority of the breakoffs are within FI-administered portions of the interview. This suggests that although breakoffs do happen, no particular questions that generate a significant number of breakoffs can be identified. Within the ACASI portion, the breakoffs are clustered in the tutorial and noncore sections, suggesting that if a breakoff does happen during the ACASI, it probably will occur right at the beginning or toward the end where respondents may be getting tired of answering questions. Respondents' reluctance to continue with a long interview also may explain why a majority of breakoffs are occurring in the FI-administered sections, which are at the beginning and the end of the entire interview. It should be noted that as a result of not being able to resolve all issues of duplicate ID numbers, and retaining all files for the breakoff analysis, the potential for double counting of interviews and instances per FI does exist.

Table 7.3 Unweighted Distribution of Breakoffs Summary

Total Audit Trail Files	68,268
Total Breakoff Instances	1,563 (2.3%)
Total Respondents	1,411
Single Instance	1,293 (91.6%)
Two Instances	92 (6.5%)
Three Instances	20 (1.4%)
Four Instances	4 (0.3%)
Five Instances	2 (0.1%)
Total Field Interviewers	523

Table 7.4 Distribution of Breakoffs, by Location in Interview

Location in Interview	***N***	**Percent of Total Breakoffs**
Total	1,563	100.0
Field Interviewer Administered	926	59.24
Respondent Interaction	543	34.74
No Respondent Interaction	383	24.50
ACASI	631	40.37
Tutorial	189	12.09
Main Core Drugs	102	6.53
Nonmedical Psychotherapeutic Drugs	52	3.33
Noncore Sections	288	15.43
Systems Crash	6	0.38

ACASI = audio computer-assisted self-interviewing.

NOTE: Respondent Interaction refers to modules administered to the respondent by the interviewer. No Respondent Interaction includes such modules as the interview debriefing.

An interesting finding is shown in ***Table 7.5*** where 90 interviewers have 5 or more instances of a breakoff and 21 interviewers have 10 or more instances, which might suggest possible interviewer problems. In this instance, the audit trail files could be used as an FI monitoring tool to alert project staff of interviewers who may have problems interacting with the CAI instrument or with respondents. Breakoff data, along with timing data, will be used in the future by NSDUH data quality staff in monitoring FI adherence to study protocols.

Table 7.5 Unweighted Distribution of Breakoffs Attributable to Field Interviewers

Number of Breakoffs	Accountable Field Interviewers (*n* = 523)	
	N	Percent
1	222	42.45
2	107	20.46
3	68	13.00
4	36	6.88
5	30	5.74
6	16	3.06
7	14	2.68
8	6	1.15
9	3	0.57
10	7	1.34
11	3	0.57
12-45	11	2.09

Tables 7.6 through ***7.8*** present the results of the analysis on the impact of using audit trail data in estimating lifetime, past year, and past month substance use prevalence. To better inform the reader of the mechanics behind the interpretation of the "Relative Percentage Change" presented in these tables, the actual unweighted sample responses for the 12 or older population for lifetime cigarette usage, along with the calculation formulas, are presented below.

Table 7.6 Relative Percentage Increase in Estimated Prevalence (Unweighted) of *Lifetime Substance Use* Using Additional Audit Trail Information

Substance	Relative Percentage Change, by Age Group			
	12+	12-17	18-49	50+
Cigarettes	1.1	3.1	0.5	0.6
Alcohol	0.9	3.5	0.2	0.4
Marijuana	1.7	3.7	1.0	5.2
Cocaine	4.1	16.2	2.9	8.3
Heroin	10.5	27.6	7.7	*
Hallucinogens	2.5	9.8	1.2	6.7
Inhalants	5.6	11.7	2.8	15.0
Pain Relievers	11.7	17.5	9.1	30.0
Tranquilizers	5.7	13.8	3.7	18.1
Stimulants	6.1	14.3	4.3	6.0
Sedatives	11.8	44.5	6.8	12.6

* Denotes low precision, no estimate reported (i.e., raw questionnaire sample <40).

As an example, ***Table 7.9*** presents the unweighted sample responses to the initial cigarette gate question: "Have you **ever** smoked part or all of a cigarette?" The respondent has the option of indicating "Yes," "No," "Don't know" or refusing.

As described earlier, the analyses took a worst case scenario, in that any indication of use was considered to be the truth. Under this assumption, the first step was to utilize information from the questionnaire and audit trail data and classify each respondent into a response category and calculate the distribution of the entire sample among these categories. "Don't know" and refusal responses are combined into the "Unknown" category.

Respondents classified into the "No" or "Unknown" categories under the raw questionnaire data could only be reclassified into the "Yes" category after accounting for audit trail information. For instance, in ***Table 7.9***, 407 respondents indicated a final "No" lifetime use of cigarettes but also had indicated a "Yes" response (as captured by their respective audit trail data at some earlier point during the interview).

Table 7.7 Relative Percentage Increase in Estimated Prevalence (Unweighted) of *Past Year Substance* Use Using Additional Audit Trail Information

	Relative Percentage Change, by Age Group			
Substance	**12+**	**12-17**	**18-49**	**50+**
Cigarettes	0.8	1.1	0.5	2.8
Alcohol	0.4	1.2	0.2	0.7
Marijuana	0.7	0.6	0.7	6.7
Cocaine	1.5	1.5	1.5	*
Heroin	3.5	6.4	2.4	*
Hallucinogens	1.8	2.4	1.6	*
Inhalants	2.5	2.3	2.7	*
Pain Relievers	1.7	2.1	1.3	16.7
Tranquilizers	1.3	1.1	1.2	7.7
Stimulants	2.4	1.8	2.7	*
Sedatives	1.3	0.9	1.5	*

*Denotes low precision, no estimate reported (i.e., raw questionnaire sample <40).

The formula for the relative percentage change is $(|(P_2 - P_1)| / P_1)*100$, where P_1 is the prevalence from the raw questionnaire data and P_2 is the prevalence when audit trail information is utilized. So, in the case of lifetime cigarette use, the relative percentage change is $(|0.6055 - 0.5992|) / 0.5992)*100 = 1.07$, rounded to 1.1, as displayed in ***Table 7.6*** for the 12 or older age category.

With regard to the lifetime prevalence measures, there is a good deal of variation across various substances and age groups. In some instances, particularly for more prevalent substances, such as cigarettes and alcohol, these slight increases in prevalence would have little or no effect. However, for some of the rarer substances, such as heroin or cocaine, even the smallest increases could have large effects. For instance, the 0.14 percentage increase in lifetime heroin use (not shown) accounts for a 10.5 percent relative increase from the estimate based on raw questionnaire data alone (see ***Table 7.6***).

Table 7.8 Relative Percentage Increase in Estimated Prevalence (Unweighted) of *Past Month Substance Use* Using Additional Audit Trail Information

	Relative Percentage Change, by Age Group			
Substance	**12+**	**12-17**	**18-49**	**50+**
Cigarettes	0.9	1.7	0.7	3.0
Alcohol	0.3	1.1	0.2	0.4
Marijuana	0.7	0.6	0.7	11.3
Cocaine	3.0	1.7	3.3	*
Heroin	7.4	*	5.3	*
Hallucinogens	2.7	1.6	3.3	*
Inhalants	3.9	3.8	4.0	*
Pain Relievers	2.3	2.5	1.6	23.1
Tranquilizers	1.9	2.0	1.6	*
Stimulants	3.6	2.2	4.3	*
Sedatives	1.9	0.0	3.1	*

*Denotes low precision, no estimate reported (i.e., raw questionnaire sample <40).

Table 7.9 Unweighted Responses to the Cigarette Gate Question among Respondents Aged 12 or Older: 2002 NSDUH Audit Trail Data

Lifetime Cigarette	Raw Questionnaire Responses		Responses Adjusted for Audit Trail Information		Relative Percentage Increase
	n	Percent	*n*	Percent	
Yes	38,233	59.92	38,640	60.55	1.07
No	25,570	40.07	25,163	39.43	1.59
Unknown	8	0.01	8	0.01	0.00

Additionally, some substance use categories (hallucinogens, inhalants, and all prescription-type psychotherapeutic drugs, or "pills") contain multiple gate questions with a last, "other, specify" question. These questions prompt the respondent to indicate any additional substances used within a particular category besides those specific substances indicated in the preceding multiple gate questions. In some instances, these questions alone account for more than a 25 percent relative increase in their respective prevalence rates. Though the "other, specify" category represents a large relative increase when considered by itself, its impact on its respective overall substance category must be calculated with regard to all the other respective gate questions before any specific conclusions can be drawn. Though not shown here, the impact of the "other, specify" category only on the overall substance category, on average, accounts for about 26 percent of the lifetime prevalence increase (high of 41.8 percent for hallucinogens and a low of 14.7 percent for inhalants). These results direct the focus of the "other, specify" analysis toward questionnaire methodology in an attempt to discover why a large number of respondents are backing up and changing their answers to these specific questions.

The relative percentage change presented in ***Tables 7.7*** and ***7.8*** follows the same premise as ***Table 7.6***, but with the focus shifting from lifetime to past year and past month use, respectively. Similar to the worst case scenario utilized for the lifetime analysis, the focus of this analysis was to capture the percentage of respondents whose audit trial information indicated a more recent period of last substance use than was indicated by their raw questionnaire data. For instance, a respondent whose raw data indicate only lifetime use may have audit trail information that shows that he or she indicated past year or past month use.

Though the prevalence analyses have shown some interesting results, they have not shown any conclusive evidence that there are any problems with the reliability of NSDUH estimates.

Management of Audit Trail Files

Some discussion should be noted on the findings concerning data management of the audit trail files. Though the size of the NSDUH study is fairly large and at first glance appears to be burdensome, it was relatively easy to manage the data. With today's advances in computer-processing abilities and the relatively low cost of storage space, all the files could be stored and processed with minimal downtime of any analyses (usually on average about 8 hours to reprocess all the 68,000 files depending on the extent of the modifications requested and about 1.5 gigabytes of storage space). This should provide some assurances to anyone who is considering performing a study of this size and wishes to utilize audit trail data.

Before using audit trail data on any large study, however, a significant change needs to be made to the processing methods used in this analysis. This entails processing the incoming

transmitted files on either a daily or at least weekly basis. Though some automated machine-editing procedures can be implemented to capture and resolve a bulk of the audit trail idiosyncrasies, it will never be able to resolve all of them. This may be relieved by implementing some automated process that will at least flag certain files that need further investigation or flag situations that indicate a problem. Some examples would be (1) identifying files with duplicate IDs and resolving them; (2) linking breakoff interviews with subsequent follow-ups; and (3) detecting instances where the FI is having technical difficulties and taking appropriate action to correct them.

Ultimately, this hands-on processing would enable project staff to immediately deflect any problems that they see occurring in the field and would produce a fairly clean data file that could be processed quickly. The authors suggest having an analyst clean and process the data as it arrives. It is the strong opinion of the authors that when the bulk of audit trail data reaches the magnitude of NSDUH and covers such a lengthy time span (i.e., 1 calendar year, divided into 4 quarters) that the data files be streamline processed (e.g., as close to real-time processing as is feasibly possible). There are several reasons for this. First, in post-processing after the end of a quarter of data collection, the data are manageable, but when dealing with the sheer number of records, detection and resolution of inconsistencies in file management become difficult. Hence, the probability of not being able to recover a complete audit trail increases the longer the duration between transmission and processing. More importantly, the closer to real-time that processing can occur, the better opportunity to utilize the data in a manner effective to resolving problems or to developing an enhancement/modification to continue to improve data quality. As an example, if a new FI is experiencing technical difficulty at the beginning of the data collection period, it would be better to rectify this situation early on instead of discovering this situation after numerous interviews have been conducted.

Future Analyses

The authors' experiences working with the 2002 NSDUH audit trail data have been enlightening. Although time and cost constraints limited the amount of analysis that could be done with the 2002 data, many potential analyses were discovered that could be performed. One further analysis would be to utilize the timing information on the initial routing through a prevalence or recency question of interest and determine the number of subsequent questions answered prior to the respondent backing up and changing his or her answers. For instance, if a respondent takes 2 seconds to respond to a lifetime substance use question with "Yes" and then after going through one or two more questions, backs up and changes that answer to "No," one might be more inclined to concede that the respondent had rushed through the initial question and did not realize his or her initial response was incorrect. Taking an opposite approach, if a respondent takes a relatively long time to answer the initial question and then proceeds through five or more questions before returning to change his or her answer, one might be more inclined to believe that the respondent is hiding the truth of his or her substance use. Furthermore, a comparison could be made between the outlying timing data of respondents who change their answers within the interview, with those who do not. This approach is still speculative, but it does provide another means to utilize additional data from the audit trail files. Based on the criteria used to determine what is a long time or a sufficient number of questions to go through prior to backing up, this approach may provide a more realistic assessment of the effect of changing answers.

Also, with advances in data-mining techniques and audit trail file data management procedures, this would be an excellent opportunity to model fraudulent cases from the audit trail data of cases already proven to be fraudulent. Though NSDUH currently conducts a random 15 percent verification of all interviews, it might be useful to calculate a predicted probability of being a fraudulent interview and specifically designate these cases to be verified within that FI's 15 percent. This method would require (1) first identifying fraudulent cases and using their audit trail data to establish response patterns and (2) continually updating the predictive model to compensate for new data. The authors plan to continue to utilize audit trail data to monitor the survey methodology and search for additional avenues of research for which audit trails would be an indispensable aspect.

References

Caspar, R. (2000, October 3-6). *Using keystroke files to identify respondent difficulties with an ACASI questionnaire*. Presented at Fifth International Conference on Social Science Methodology, Cologne, Germany.

Caspar, R., & Couper, M. (1997). Using keystroke files to assess respondent difficulties with an audio-CASI instrument. In *Proceedings of the American Statistical Association, Survey Research Methods Section* (pp. 239-244). Alexandria, VA: American Statistical Association.

Lessler, J. T., Caspar, R. A., Penne, M. A., & Barker, P. R. (2000). Developing computer assisted interviewing (CAI) for the National Household Survey on Drug Abuse. *Journal of Drug Issues, 30*(1), 9-34.

Office of Applied Studies. (2001). *Development of computer-assisted interviewing procedures for the National Household Survey on Drug Abuse* (DHHS Publication No. SMA 01-3514, Methodology Series M-3). Rockville, MD: Substance Abuse and Mental Health Services Administration. [Available at http://www.oas.samhsa.gov/nhsda/CompAssistInterview/toc.htm]

Office of Applied Studies. (2002). *Results from the 2001 National Household Survey on Drug Abuse: Volume I. Summary of national findings* (DHHS Publication No. SMA 02-3758, NHSDA Series H-17). Rockville, MD: Substance Abuse and Mental Health Services Administration. [Available at http://www.oas.samhsa.gov/p0000016.htm#Standard]

Penne, M., Snodgrass, J., & Barker, P. (2002, November 14-17). *Analyzing audit trails in the National Survey on Drug Use and Health: Means for maintaining and improving data quality*. Presented at the International Conference on Questionnaire Development, Evaluation, and Testing Methods, Charleston, SC.

8. Evaluation of Follow-Up Probes to Reduce Item Nonresponse in NSDUH

Rachel A. Caspar, Michael A. Penne, and Elizabeth Dean
RTI International

Introduction

Technological advances in survey research over the past 15 years offer the ability to improve the quality of face-to-face survey data in part by reducing item nonresponse. Assuming the instrument has been programmed correctly, computer-assisted interviewing (CAI) can eliminate missing data caused by confusing skip instructions, hard-to-locate answer spaces, and simple inattention to the task at hand. Yet in and of itself, CAI cannot reduce the item nonresponse created when a respondent chooses to give a "Don't know" or "Refused" response to a survey question. Previous research (see, e.g., Turner, Lessler, & Gfroerer, 1992) has shown that these types of responses are more common in self-administered questionnaires than in those administered by an interviewer. Most likely, this occurs because in a self-administered interview the interviewer does not see the respondent's answer and thus cannot probe or follow up on these types of incomplete responses. This chapter introduces a methodology designed to reduce item nonresponse to critical items in the audio computer-assisted self-interviewing (ACASI) portion of the questionnaire used in the National Survey on Drug Use and Health (NSDUH).[1] Respondents providing "Don't know" or "Refused" responses to items designated as essential to the study's objectives received tailored follow-up questions designed to simulate interviewer probes.

Causes of Item Nonresponse

Item nonresponse occurs in surveys when respondents answer "Don't know" or "Refused" to a question, or when respondents leave an item blank. Item nonresponse may be attributed to four types of causes: mode of interview, the questionnaire, the respondent, and the interviewer (de Leeuw, 1999). The effect of interview mode on item nonresponse is most dramatic in self-administered questionnaires when an interviewer is not present to follow up with a respondent by probing a refusal, a vague or incomplete answer, or a "Don't know" response. Additionally, paper questionnaires present respondents with the challenge of determining where and how to record their responses (Jenkins & Dillman, 1997). In contrast, interviewer-administered questionnaires may be less susceptible to item nonresponse caused by poor formatting, but are subject to item nonresponse due to question sensitivity. People are less willing to report sensitive behaviors when they have to speak them out loud to another person (Rogers, Miller, Forsyth, Smith, &

[1] Prior to 2002, the survey was called the National Household Survey on Drug Abuse (NHSDA). In this report, it is referred to as NSDUH, regardless of the survey year.

Turner, 1996; Sudman, Bradburn, Blair, & Stocking, 1977) than when they can privately enter their response.

Questionnaire design has a dramatic impact on item nonresponse in self-administered questionnaires. Self-administered questionnaires provide privacy and tend to reduce the effects of sensitivity on item nonresponse. But if a respondent does not realize that he or she even needs to answer a particular question because the layout is confusing, or the font is difficult to read, self-administration is of little help (de Leeuw, 1999). The same problems of missing a skip pattern or not recording an answer in the right place also can plague interviewer-administered questionnaires (Alwin, 1977). Additionally, poorly worded questions can confuse respondents to the point where they do not know how to answer with anything other than a "Don't know."

Respondent characteristics also are a factor in item nonresponse. Intensive cognitive demands and lengthy interviews can fatigue respondents, causing them to skip important questions in self-administered paper interviews or to provide a lot of "Don't know" responses. Respondents may become angry with the entire process and refuse to answer items. In other situations, respondents who are poorly educated or have reading or comprehension difficulties may be more likely to skip items (Craig & McCann, 1978; Groves, 1989). In addition, respondents may refuse because they object to providing sensitive information, even when an interviewer is not present.

Finally, the interviewers themselves can contribute to item nonresponse. Interviewers who are poorly trained or who are trying to move through the interview too quickly may accept a "Don't know" or "Refused" response without probing or otherwise trying to persuade the respondent to provide a valid answer. For open-ended questions, they may accept a vague answer that is meaningless in the context of the analysis objectives (de Leeuw 1999). Interviewers also can have a negative effect when they come to the project with preconceptions about the data being collected. Studies have shown that item nonresponse is greater when interviewers do not believe people will be willing to answer the questions. In contrast, interviewers who are confident in their ability to obtain complete data from respondents and who are comfortable with the subject matter of the study generally obtain lower rates of item nonresponse (Bailar, Bailey, & Stevens, 1977; Groves, 1989; Lessler & Kalsbeek, 1992; Sudman et al., 1977).

Methods Used to Reduce the Impact of Item Nonresponse

Many approaches have been developed to reduce the effects of item nonresponse in survey interviews. Lessler and Kalsbeek (1992) distinguished between preventive measures, which reduce the likelihood that item nonresponse will occur, and post hoc imputation methods, which adjust existing data to account for items that are missed. Preventive measures include a wide range of strategies, including the following:

- easing the reporting of sensitive data by using a self-administered interview mode or by introducing randomized response techniques to an interviewer-administered mode;
- pretesting the questionnaire to ensure that the response task is as simple as possible;

- designing paper forms that are easy to follow;
- training interviewers to probe effectively the "Don't know" and "Refused" responses to elicit a usable answer;
- selecting confident interviewers who believe that respondents will be interested in answering questions about the survey's topics;
- tailoring the instrument to the needs of respondents who may be very young, elderly, or disabled; and
- editing data through comparisons with other sources.

Preventive measures taken to reduce item nonresponse should be determined by the particular factors that are likely to affect item nonresponse for a specific study. In some situations, the less expensive approaches, such as revising the questionnaire formatting or improving interviewer training, may go a long way. In situations where self-administration is desirable, such as when collecting highly sensitive data, conversion to a computer-assisted mode of administration may be warranted.

Despite the positive effects of using preventive measures to reduce item nonresponse and of using reliable imputation techniques to obtain analyzable data, item nonresponse remains an especially serious concern in surveys measuring rare and sensitive behaviors. The low incidence of these types of behaviors, coupled with respondents' reluctance to report them, means that even a small amount of item nonresponse has the potential to significantly affect data quality. The higher the rate of item nonresponse for a variable, the greater the concern that the estimates derived from that variable may be biased.

The remainder of this chapter discusses a preventive methodology for reducing item nonresponse in NSDUH, a study that collects sensitive data on substance use in the United States via a self-administered questionnaire. Until 1999, the basic data collection methodology of the survey had remained unchanged since its inception in 1971. In-person interviews were conducted, with some questions administered by the interviewer and the more sensitive data collected in a self-administered format. Questions in the interviewer-administered sections were mostly demographic in nature. The self-administered questions were given to the respondent on "answer sheets," which he or she read and filled out without assistance from the interviewer, unless requested. The answer sheet methodology was used to encourage honest reporting of sensitive information by allowing respondents to report these behaviors in privacy. After its completion, each answer sheet was placed in an envelope, and at the end of the interview the envelope was sealed and mailed to the contractor's data processing site. Thus, there was no field editing of the respondent-completed answer sheets, and respondents were not recontacted to resolve inconsistencies. To further protect the respondent's privacy and further encourage honest reporting, no respondent names were ever recorded.

Beginning in 1996, research into the feasibility and potential benefits of converting NSDUH to a CAI format was undertaken. Research showing the feasibility of ACASI and its potential for improving the reporting of sensitive behaviors was pivotal in the decision made by the Substance Abuse and Mental Health Services Administration (SAMHSA) (Duffer, Lessler, Weeks, & Mosher, 1996; Turner, Ku, Sonenstein, & Pleck, 1996).

ACASI methodology allows the respondent to listen to questions through a headset and/or to read the questions on the computer screen. Respondents also key their own answers into the computer while the interviewer remains out of viewing range of the screen. Use of ACASI maintains the self-administered format already found to increase reporting of sensitive behaviors, and greater privacy can be ensured for the respondent even in interview settings that might not otherwise be considered sufficiently private. In addition, respondents with limited reading abilities can be accommodated with no loss of privacy. Programming the questionnaire also allows for more complex skip logic in a format where the routing is less visible to the respondent. Thus, it is easier for a respondent to "navigate" the instrument because the question sequence is automatic. It also may be less obvious to the respondent how answering a question in a particular way will influence the number and type of additional questions asked. In summary, ACASI meets the need for privacy, the need to reduce respondent burden, and the desire for improved data quality, both by encouraging more honest reporting and by using edit checks at the time of interview.

A Methodology for Reducing Item Nonresponse in NSDUH

When the survey was administered as a paper-and-pencil instrument, interviewers did not instruct respondents on how to record a "Don't know" or "Refused" response unless explicitly asked. Questions left blank on the paper instrument were difficult to categorize for this reason. In the 1999 survey, questions left blank accounted for an average of 16 percent across a set of 26 key items.[2] It was impossible to tell whether a blank question meant the respondent did not know the answer, did not want to answer the question, or had skipped the question inadvertently. With the transition to ACASI, the decision was made to require the respondent to enter a response to each question, thus eliminating any unanswered questions. As a result, however, it was necessary to provide respondents with specific instructions on how to enter a "Don't know" or "Refused" response.[3]

As with any self-administered questionnaire, the quality of ACASI data may suffer because an interviewer cannot probe to follow up on "Don't know" or "Refused" responses. During a face-to-face interview, the interviewer typically is trained to follow up "Don't know" responses with neutral probes to try to get the respondent to choose an answer that comes closest to his or her situation, provide a best guess or estimate, or spend a little more time thinking about the question in case that helps him or her to recall the information. Similarly, when a respondent refuses to answer a question, the interviewer is trained in procedures for encouraging the respondent to answer the question before assigning the refusal code and moving to the next question.

[2] The 26 key items include *lifetime* use questions for cigarettes, smokeless tobacco, alcohol, marijuana, cocaine, crack, heroin, LSD, and PCP; *recency* of use questions for cigarettes, smokeless tobacco, alcohol, any hallucinogen, any inhalant, and nonprescription use of prescription pain relievers, prescription tranquilizers, prescription stimulants, and prescription sedatives; and *30-day frequency* questions for use of cigarettes, alcohol, marijuana, cocaine, crack, heroin, any hallucinogen, and any inhalant.

[3] Two of the computer's function keys are used to record these responses. F3 is used for "Don't know," and F4 is used for "Refused." The interviewer points out the location of these keys when he or she teaches the respondent to use the laptop computer.

In order to continually identify ways to improve data quality, an effort was made to minimize the "Don't know" and "Refused" answers in the 2000 survey. To do this, the item nonresponse rates for the 1999 survey were first reviewed. Attention was focused on a set of items in the instrument that SAMHSA staff identified as critical to their reporting needs. These are items that have been included in the survey each year and comprise an important component of the trend data that SAMHSA reports. In many cases, they also are items that serve as gate questions for additional questions about the respondent's drug use. Without a substantive response to these items, additional follow-up questions are not possible. These items ask the following:

- whether the respondent has ever used a particular substance (used to generate lifetime use estimates),
- the last time the respondent used the substance (used to generate recency of use estimates), and
- the number of days the respondent has used the substance in the past 30 days (used to generate 30-day frequency of use estimates).

Many of these items also are integral to the determination of whether a case is defined as usable for analysis or not. To be defined as usable (and thus included on the analysis data file), a case was required to have both of the following:

- either a "yes" or "no" response to the lifetime use of cigarettes question; and
- a "yes" or "no" response to at least nine of the following additional lifetime use questions: chewing tobacco, snuff, cigars, alcohol, marijuana, cocaine (in any form), heroin, hallucinogens, inhalants, pain relievers, tranquilizers, stimulants, and sedatives.

Using follow-up questions for all questions in the ACASI interview was not considered for fear that respondents would become annoyed if the items appeared too frequently and terminate the interview. Also, a conscious effort was made not to increase significantly the overall length of the interview and respondent burden.

Magnitude of Item Nonresponse for Critical Items

A review of the 1999 data showed that item nonresponse for the lifetime use items was generally quite low. Refusals ranged from a low of 0.15 percent for pipe tobacco to a high of 0.50 percent for marijuana. The average refusal rate for lifetime use items was 0.25 percent. "Don't know" responses were even less common, ranging from 0.05 percent for cigarettes to 0.14 percent for phencyclidine (PCP). The average rate of "Don't know" responses to the lifetime use items was 0.08 percent.

The item used to assess recency of use is asked only of those respondents who have reported lifetime use of a given drug. In the tobacco module (the module that includes questions on cigarettes, chewing tobacco, snuff, and cigars), respondents are asked to report the recency of their use by selecting one of three categories:

- more than 30 days ago but within the past 12 months;

- more than 12 months ago, but within the past 3 years; or
- more than 3 years ago.

For the other drugs (alcohol, marijuana, cocaine, crack, heroin, hallucinogens, inhalants, and nonmedical use of prescription pain relievers, tranquilizers, sedatives, or stimulants), recency of use is measured with three slightly different categories:

- within the past 30 days;
- more than 30 days ago but within the past 12 months; or
- more than 12 months ago.

In contrast to the lifetime use items, nonresponse for the recency of use items is more attributable to "Don't know" responses than to refusals. Among respondents who reported lifetime use and thus were routed to the recency question, rates of "Don't know" responses ranged from 0.42 percent for cocaine to 3.97 percent for prescription pain relievers, with a mean rate of 1.47 percent. Refusal rates ranged from 0.11 percent for cigars to 1.82 percent for prescription pain relievers, with a mean rate of 0.79 percent. These results suggest that rather than finding this information overly sensitive, respondents may simply have difficulty recalling it.

Finally, respondents who reported use during the past 30 days were asked to report the number of days they used the substance during that time period (referred to as 30-day frequency).[4] As with the recency items, "Don't know" responses were more common than refusals. Rates of "Don't know" responses ranged from 0.0 percent for heroin to 4.55 percent for crack, with a mean rate of 2.95 percent. Refusal rates ranged from 0.0 percent for crack to 1.45 percent for heroin. The mean rate of refusal was 0.61 percent for the 30-day frequency item.

Taken in total, the good news from this review of the 1999 data is that item nonresponse rates for these critical items were low for the most part. However, the prevalence of some of the behaviors asked about in NSDUH is quite small in the household population. For example, based on the 1999 data, prevalence of heroin use in the past 30 days was estimated at only 0.1 percent for the national sample (Office of Applied Studies [OAS], 2000). Therefore, even a small item nonresponse rate has the potential to seriously affect the quality of the population estimates.

Developing the Follow-Up Items for the 2000 Survey

As noted above, item nonresponse for the critical set of items was generally low. However, in an effort to further improve the quality of the data, RTI and SAMHSA staff decided to investigate the use of an ACASI procedure that could mimic the role the interviewer plays in probing for more complete data when a respondent provides an initial "Don't know" or "Refused" response in an interviewer-administered survey.

[4] This item is not asked for the four different types of prescription pills.

Development work began by first considering why a respondent might decide to enter a "Don't know" or "Refused" response to one of the critical items. Hypotheses as to why each item might cause a respondent to give either a "Don't know" or a "Refused" response were considered separately for each of the three sets of critical items.

Refusals. It seemed that the reason for a refusal would be the same across each of the three types of critical items. Namely, the hypothesis was that respondents would refuse to answer a given question if they were concerned about the privacy of their responses or if they questioned the necessity of collecting information of such a personal nature. A third reason for refusing that is commonly discussed in the literature (e.g., de Leeuw, 1992; DeMaio, 1984) is the respondent's desire to "save face" or appear "desirable" to the interviewer. However, given the ACASI structure of the interview, this did not seem to be a likely reason for refusals in the 2000 questionnaire. To alleviate respondent concerns, the following question text was drafted (the example shown here would appear if a respondent refused to answer the question about lifetime use of heroin):

> The information respondents provide about their drug use is very important to the success of this study. We recognize that this information is personal. Please remember that the answers you give will be kept strictly confidential and they will never be linked to your name:
>
> Please reconsider answering this question: Have you **ever**, even once, used heroin?

This text was designed to mirror what an interviewer might say in a situation where a respondent initially refused to answer a question of this type.

Don't Knows. In contrast, it was hypothesized that respondents would have differing reasons for providing a "Don't know" response to each of the three categories of key items. Answering "Don't know" to a lifetime use item could be a way to avoid answering a sensitive question (a tacit refusal). However, it also could be given if a respondent truly did not understand the drug terminology used in the question or if the respondent was not sure if the drug he or she had used should be included in a particular class of drugs. It seemed ill-advised to try to "force" these latter types of respondents who were providing legitimate "Don't know" responses to give a substantive response. For this reason, a follow-up question for "Don't know" responses was not included for the lifetime use questions.

Next, possible reasons why a respondent would give a "Don't know" response to a recency question were considered. As noted earlier, the recency questions in NSDUH ask respondents to select one of three categories to indicate how long it has been since their last use of a drug. It seemed likely that a reason for providing a "Don't know" response would be uncertainty about which category to choose. Respondents whose most recent use occurred at the "seams" where the categories meet might have difficulty deciding between response categories and opt to give a "Don't know" response instead. The following question was drafted to follow up on "Don't know" responses to the recency questions (again using heroin as an example):

> What is your **best guess** of how long it has been since you **last** used heroin?

By including the phrase "best guess," the hope was that respondents would be willing to provide a substantive answer even if they were not absolutely sure it was correct. For these critical items, it was decided even a best guess was preferable to a "Don't know" response.

Similar logic was used in developing a follow-up question for "Don't know" responses to the 30-day frequency of use questions. These questions ask the respondent to report the total number of days during the past 30 days that he or she used the particular drug. Respondents are instructed to key their answer in directly rather than select a category. Any response up to 30 is allowed. It was theorized that a "Don't know" response could indicate the respondent was unsure of the exact number of days that he or she had used the drug and was choosing to give a "Don't know" response rather than guess. For this reason, a follow-up question similar to that prepared for the recency items above was drafted:

> What is your **best guess** of the number of days you used heroin during the past 30 days?

In addition, however, response categories were included instead of leaving the question open-ended with the goal of further simplifying the respondent's task. The response categories created were as follows:

- 1 or 2 days,
- 3 to 5 days,
- 6 to 9 days,
- 10 to 19 days,
- 20 to 29 days, and
- all 30 days.

In summary, these follow-up questions were expected to be effective because they were tailored to the particular kind of item nonresponse the respondent expressed—either a "Don't know" or a "Refused" answer. Only refusal responses were followed up for the lifetime use question. For the recency and 30-day frequency items, follow-up questions were included for both "Don't know" and "Refused" responses.

Results

The analytic results reported in this section are based on data collected for the 2000 survey (*n* = 71,764) and are unweighted.[5] In total, 2,122 respondents (3.0 percent) triggered at least one of the 38 item nonresponse follow-up questions in 2000.

[5] The decision was made to use unweighted data because of the interest in describing the characteristics of respondents who triggered follow-up items and those who went on to provide a substantive response. Later in the chapter, both weighted and unweighted analyses are presented when the impact of the follow-up items on prevalence rates is addressed.

Descriptive Analyses

Table 8.1 provides the demographic characteristics of those respondents who triggered at least one follow-up item.

Table 8.1 Demographic Distribution of Respondents Triggering at Least One Follow-Up Question

Demographic	Triggering Sample	Total Respondent Sample	Percent
Total	2,122	71,764	3.0
Age Group (Years)			
12-17	986	25,717	3.8
18-25	621	22,613	2.8
26-49	354	16,711	2.1
50+	161	6,724	2.4
Gender			
Male	1,165	34,386	3.4
Female	957	37,378	2.6
Race/Ethnicity			
Hispanic	262	9,393	2.8
Non-Hispanic Black	315	8,721	3.6
Non-Hispanic Non-Black	1,545	53,650	2.9
Education (18+ Only)			
Less Than High School	296	8,376	3.5
High School Graduate	450	16,026	2.8
More Than High School	390	21,265	1.8

Lifetime Use Follow-Up. Across all eight drug types (cigarettes, alcohol, marijuana, cocaine including crack, crack, heroin, lysergic acid diethylamide [LSD], and PCP), the lifetime use follow-up question was triggered only 499 times by a total of 332 respondents (***Table 8.2***). Multiple lifetime use follow-up items were triggered by 80 respondents (24.1 percent). The maximum number of lifetime use follow-up items triggered by a single respondent was five.

Table 8.2 Responses to the Lifetime Use Follow-Up Items (Summarized Across All Drug Types)

Response Given to Follow-Up	Percent	*N*
Yes	25.7	128
No	14.0	70
Refused	59.1	295
Don't Know	1.2	6
Total	100.0	499

The follow-up item was not completely effective in motivating respondents to provide a substantive response to the lifetime use question. More than half the respondents (59 percent) simply refused to answer the follow-up question just as they had refused the original item. A substantive response was obtained from only 40 percent of these respondents (see ***Table 8.2***). ***Table 8.2*** also shows the full breakdown of responses for the lifetime use follow-up item. Interestingly, among those respondents who did provide a substantive response to a lifetime use follow-up item (n = 198), nearly two thirds (65 percent) reported lifetime use. This compares with an unweighted average over these measures of 25 percent for those who responded to the original lifetime use questions. And it suggests that disclosure concerns may explain at least some respondents' reluctance to answer the initial questions.

The lifetime use follow-up item was more effective in persuading younger respondents to provide a substantive response. ***Table 8.3*** shows that more than half of the youth respondents (60 percent) who triggered a lifetime use follow-up item provided a substantive response. That percentage decreased with increasing age; only 26 percent of respondents aged 50 or older provided a substantive response to the follow-up.

Table 8.3 Age of Respondents Converted by the Follow-Up Item for Lifetime Use (Summarized Across All Drug Types)

Respondent's Age Group	Number of Follow-Up Respondents	Respondents Converted	
		Percent	*N*
12 to 17	117	59.8	70
18 to 25	160	40.6	65
26 to 49	169	29.0	49
50 or Older	53	26.4	14

Results varied somewhat by specific drug type. ***Table 8.4*** provides results for each drug for the total group of respondents who triggered lifetime follow-up items and also by age group. Looking first at the results for all lifetime follow-up respondents, ***Table 8.4*** suggests that all respondents who triggered the lifetime follow-up for cigarettes (n = 29) provided a substantive response. However, this interpretation is somewhat incorrect. As a result of the usable case rules developed for the 2000 survey (described later in this chapter), respondents with missing data for the cigarette lifetime use question were designated unusable and dropped from the final analysis file.[6] A more appropriate interpretation of the lifetime follow-up item for cigarettes is that at least 29 cases that would otherwise have been designated unusable were retained.

A majority of respondents provided substantive responses to the lifetime follow-up for alcohol. However, for the remainder of the drugs, only a minority of respondents provided substantive answers. For these drugs, most respondents either again entered a "Refused" response or selected the "Don't know" option. As these drugs are illegal for all persons, one can assume that providing information about use of these drugs would be considered more sensitive than for cigarettes or alcohol. This may be the reason that fewer respondents were converted by the follow-up item. It is interesting to note, however, that for nearly all the drugs, among respondents who provided a substantive answer, there were more affirmative responses than negative responses.

Subgroup results must be interpreted cautiously due to small sample sizes. However, for most drugs, youth respondents were more likely to provide substantive responses than were older respondents. Most notably, more than half (59 percent) of the youth respondents who triggered the marijuana lifetime follow-up item reported use or nonuse of the drug. This rate decreased to 49 percent for the 18 to 25 age group, 43 percent for the 26 to 49 age group, and only 13 percent of the respondents aged 50 or older who reported marijuana use or nonuse for the lifetime use follow-up item.

Recency of Use Follow-Up. As noted earlier, both "Don't know" and "Refused" responses to any of the original recency items triggered a follow-up. A small group of these respondents (14.4 percent) triggered more than one recency follow-up. A total of 1,184 respondents triggered at least one recency follow-up item. A total of 1,425 recency follow-up items were triggered across all drug types. The greatest number of recency follow-ups triggered by a single respondent was nine. Respondents were more likely to trigger the follow-up item by providing a "Don't know"

[6] In the 2000 survey, 65 cases were defined as unusable because a valid response was not obtained for the cigarette lifetime use question or its associated follow-up question.

Table 8.4 Answers Provided to the Follow-Up Questions for Lifetime Use, by Age of Respondent

Response to Lifetime Use Follow-Up	Cigarettes		Alcohol		Marijuana		Cocaine (Including Crack)		Crack		Heroin		LSD		PCP	
	Percent	*N*	Percent	*N*	Percent	*N*	Percent	*N*	Percent	*N*	Percent	*N*	Percent	*N*	Percent	*N*
Total	100.0	29	100.0	47	100.0	192	100.0	75	100.0	7	100.0	29	100.0	50	100.0	38
Yes	48.3	14	34.0	16	37.0	71	17.3	13	42.9	3	3.4	1	10.0	5	2.6	1
No	51.7	15	31.9	15	8.3	16	9.3	7	–	0	3.4	1	6.0	3	13.2	5
Refused	–	0	34.1	16	54.7	105	70.7	53	57.1	4	86.2	25	82.0	41	81.6	31
Don't Know	–	0	–	0	–	0	2.7	2	–	0	7.0	2	2.0	1	2.6	1
12 to 17	100.0	22	100.0	20	100.0	34	100.0	6	100.0	2	100.0	6	100.0	6	100.0	8
Yes	45.5	10	30.0	6	47.1	16	33.3	2	–	0	–	0	16.7	1	12.5	1
No	54.6	12	40.0	8	11.8	4	–	0	–	0	–	0	16.7	1	37.5	3
Refused	–	0	30.0	6	41.2	14	66.7	4	100.0	2	100.0	6	50.0	3	37.5	3
Don't Know	–	0	–	0	–	0	–	0	–	0	–	0	16.6	1	12.5	1
18 to 25	100.0	1	100.0	14	100.0	65	100.0	31	100.0	1	100.0	10	100.0	17	100.0	13
Yes	100.0	1	42.8	6	43.1	28	25.8	8	100.0	1	–	0	5.9	1	–	0
No	–	0	28.6	4	6.1	4	9.7	3	–	0	10.0	1	5.9	1	15.4	2
Refused	–	0	28.6	4	50.8	33	61.3	19	–	0	90.0	9	88.2	15	84.6	11
Don't Know	–	0	–	0	–	0	3.2	1	–	0	–	0	–	0	–	0
26 to 49	100.0	3	100.0	7	100.0	77	100.0	29	100.0	3	100.0	9	100.0	22	100.0	12
Yes	66.7	2	28.6	2	32.5	25	3.5	1	33.3	1	11.1	1	4.5	1	–	0
No	33.3	1	28.6	2	10.4	8	10.3	3	–	0	–	0	9.1	2	–	0
Refused	–	0	42.8	3	57.1	44	87.7	24	66.7	2	66.7	6	86.4	19	100.0	12
Don't Know	–	0	–	0	–	0	3.5	1	–	0	22.2	2	–	0	–	0
50 or Older	100.0	3	100.0	6	100.0	16	100.0	9	100.0	1	100.0	4	100.0	5	100.0	5
Yes	33.3	1	33.3	2	12.5	2	22.2	2	100.0	1	–	0	20.0	1	–	0
No	66.7	2	16.7	1	–	0	11.1	1	–	0	–	0	–	0	–	0
Refused	–	0	50.0	3	87.5	14	66.7	6	–	0	100.0	4	80.0	4	100.0	5
Don't Know	–	0	–	0	–	0	–	0	–	0	–	0	–	0	–	0

response to the original recency item. In nearly two thirds (65 percent) of the cases, a "Don't know" response served as the trigger for the follow-up item.

Results from the recency follow-up items are included in ***Tables 8.5*** and ***8.6***. The results for the cigarette recency item are shown separately (in ***Table 8.5***) due to a slight difference in the response categories used in the tobacco module. The data in ***Table 8.5*** show that 129 respondents triggered the cigarette recency follow-up item. Slightly more than half of these respondents (52 percent) provided a substantive response to the item. Perhaps not surprisingly, respondents who refused to answer the original recency item were significantly (p <0.05) less likely to provide a substantive answer to the follow-up than were respondents who originally provided a "Don't know" response (34 vs. 58 percent). Respondents who did provide a substantive response to the recency follow-up were most likely to report use in the distant past (more than 3 years ago).

Table 8.5 Responses to the Recency Follow-Up Item for Cigarettes

Response Given to Follow-Up Item	Overall		"Don't Know"		"Refused"	
	Percent	*N*	Percent	*N*	Percent	*N*
More Than 30 Days Ago but within Past Year	8.5	11	8.3	8	9.4	3
1 to 3 Years Ago	11.6	15	12.4	12	9.4	3
More Than 3 Years Ago	31.8	41	37.1	36	15.6	5
Refused	14.7	19	5.1	5	43.7	14
Don't Know	33.4	43	37.1	36	21.9	7
Total	100.0	129	100.0	97	100.0	32

Table 8.6 Responses to the Recency Follow-Up Items
(Summarized Across All Drug Types, Excluding Tobacco)

Response Given to Follow-Up Item	Overall		"Don't Know"		"Refused"	
	Percent	*N*	Percent	*N*	Percent	*N*
Past 30 Days	8.8	100	6.6	47	12.5	53
More Than 30 Days Ago but within Past Year	10.2	116	13.4	95	5.0	21
More Than a Year Ago	20.1	228	22.4	159	16.3	69
Refused	23.8	270	4.5	32	56.1	238
Don't Know	37.1	421	53.2	378	10.1	43
Total	100.0	1,135	100.0	711	100.0	471

As shown in ***Table 8.6***, the recency follow-up was less successful in converting respondents who encountered the item in other drug modules. The follow-up was included in 11 other drug modules for alcohol, marijuana, cocaine including crack, crack, heroin, hallucinogens, inhalants, pain relievers, tranquilizers, stimulants, and sedatives. In only 40 percent of the situations where a recency follow-up was triggered for 1 of these 11 drugs did respondents provide a substantive response. As in ***Table 8.5***, respondents who triggered the recency follow-up by providing a "Refused" response to the original item were less likely to provide a substantive answer to the follow-up than were respondents who originally provided a "Don't know" response (34 vs. 42 percent).[7] Respondents who did provide a substantive response to the recency follow-up were most likely to report use in the distant past (more than a year ago).

[7] Statistical significance of the results in ***Table 8.6*** was not tested due to the lack of an independent sample—the same respondents could have triggered more than one recency follow-up across the different drug types. Small sample sizes precluded conducting drug-specific significance tests as was done for cigarettes in ***Table 8.5.***

In contrast to the lifetime use follow-up item, the recency follow-up was not consistently more effective in persuading younger respondents to provide a substantive response. ***Table 8.7*** shows that youths who triggered the recency follow-up were more likely than the older respondents to provide a substantive follow-up for the cigarette recency item. However, for the remaining recency follow-up items (***Table 8.8***), the results are less clear, with the only consistent pattern that the oldest age group (50 or older) was least likely to provide a substantive response to the follow-up.

Table 8.7 Respondents Converted by the Recency Follow-Up Item for Cigarettes, by Age

Respondent Age Group	Overall			"Don't Know" Only			"Refused" Only		
	Number of Follow-Ups Triggered	Percent Con-verted	*N*	Number of Follow-Ups Triggered	Percent Con-verted	*N*	Number of Follow-Ups Triggered	Percent Con-verted	*N*
12 to 17	61	55.7	34	42	64.3	27	19	36.8	7
18 to 25	28	53.6	15	20	55.0	11	8	50.0	4
26 to 49	15	46.7	7	12	58.3	7	3	–	0
50 or Older	25	44.0	11	23	47.8	11	2	–	0
Total	129	51.9	67	97	57.7	56	32	34.4	11

30-Day Frequency Follow-Up. Unlike the recency follow-up, the 30-day frequency follow-up presented the same "best guess" question to respondents answering both "Don't know" and "Refused." The expectation was that respondents who had already admitted to using a drug, and using it within the past 30 days, would be comfortable reporting the number of days they used the drug. This turned out to be the case. In about 82 percent of instances where the follow-up was asked, it was triggered by a "Don't know" answer. Thirty-day frequency follow-ups were included for cigarettes, chewing tobacco, snuff, pipe tobacco, alcohol, marijuana, cocaine including crack, crack, heroin, hallucinogens, and inhalants.

Across all drug types, the 30-day frequency follow-up was triggered 1,006 times by a total of 846 respondents. A total of 135 respondents (16 percent) triggered more than one 30-day follow-up item. The greatest number of 30-day follow-ups triggered by an individual respondent was four. In contrast to the recency and lifetime follow-ups, the 30-day frequency follow-up was very successful. In about 80 percent of cases where the respondent answered "Don't know" or "Refused" to a 30-day frequency item, the follow-up item yielded a substantive response, that is, a drug use frequency (see ***Table 8.9***). ***Table 8.9*** also shows that respondents who initially refused were far less likely to provide a substantive answer than respondents who initially provided a "Don't know" answer (54 vs. 85 percent).[8] However, the 30-day frequency follow-up was far more successful than the lifetime and recency follow-ups, even for respondents who had initially refused to answer the item.

[8] Statistical significance of the results in ***Table 8.9*** was not tested due to the lack of an independent sample—the same respondents could have triggered more than one recency follow-up across the different drug types. Small sample sizes precluded conducting drug-specific significance tests as was done for cigarettes in ***Table 8.5***.

Table 8.8 Respondents Converted to a Substantive Response, by the Recency Follow-Up Item

Age of Respondent	Alcohol						Marijuana					
	Total Triggered	Percent Converted	Triggered from "Don't Know"	Percent Converted	Triggered from "Refused"	Percent Converted	Total Triggered	Percent Converted	Triggered from "Don't Know"	Percent Converted	Triggered from "Refused"	Percent Converted
12 to 17	137	48.9	101	53.5	36	36.1	49	49.0	27	51.9	22	45.5
18 to 25	37	45.9	23	52.2	14	35.7	70	55.7	21	52.4	49	57.1
26 to 49	19	63.2	13	76.9	6	33.3	36	52.8	14	50.0	22	54.5
50 or Older	19	47.4	13	53.8	6	33.3	7	42.9	2	100.0	5	20.0
Total	212	49.5	150	55.3	62	35.5	162	52.5	64	53.1	98	52.0

Age of Respondent	Cocaine (Including Crack)						Crack					
	Total Triggered	Percent Converted	Triggered from "Don't Know"	Percent Converted	Triggered from "Refused"	Percent Converted	Total Triggered	Percent Converted	Triggered from "Don't Know"	Percent Converted	Triggered from "Refused"	Percent Converted
12 to 17	14	42.9	5	20.0	9	55.6	2	–	1	–	1	–
18 to 25	16	31.2	6	–	10	50.0	8	50.0	5	80.0	3	–
26 to 49	8	50.0	6	66.7	2	–	3	66.7	2	100.0	1	–
50 or Older	3	33.3	2	50.0	1	–	1	–	1	–	0	–
Total	41	39.0	19	31.6	22	45.5	14	42.8	9	66.7	5	–

Age of Respondent	Heroin						Hallucinogens					
	Total Triggered	Percent Converted	Triggered from "Don't Know"	Percent Converted	Triggered from "Refused"	Percent Converted	Total Triggered	Percent Converted	Triggered from "Don't Know"	Percent Converted	Triggered from "Refused"	Percent Converted
12 to 17	4	50.0	3	66.7	1	–	41	36.6	27	40.7	14	28.6
18 to 25	2	–	1	–	1	–	31	38.7	17	47.1	14	28.6
26 to 49	2	–	0	–	2	–	8	50.0	5	80.0	3	–
50 or Older	1	–	1	–	0	–	3	–	0	–	3	–
Total	9	22.2	5	40.0	4	–	83	37.4	49	46.9	34	23.5

Table 8.8 Respondents Converted to a Substantive Response, by the Recency Follow-Up Item (continued)

Age of Respondent	Inhalants						Pain Relievers					
	Total Triggered	Percent Converted	Triggered from "Don't Know"	Percent Converted	Triggered from "Refused"	Percent Converted	Total Triggered	Percent Converted	Triggered from "Don't Know"	Percent Converted	Triggered from "Refused"	Percent Converted
12 to 17	125	25.6	78	28.2	47	21.3	139	36.0	111	37.8	28	28.6
18 to 25	24	37.5	11	45.5	13	30.8	68	33.8	44	38.6	24	25.0
26 to 49	16	50.0	9	44.4	7	57.1	46	28.3	25	32.0	21	23.8
50 or Older	3	33.3	2	50.0	1	–	21	9.5	11	–	10	20.0
Total	168	29.8	100	32.0	68	26.5	274	32.1	191	35.1	83	25.3

Age of Respondent	Tranquilizers						Stimulants					
	Total Triggered	Percent Converted	Triggered from "Don't Know"	Percent Converted	Triggered from "Refused"	Percent Converted	Total Triggered	Percent Converted	Triggered from "Don't Know"	Percent Converted	Triggered from "Refused"	Percent Converted
12 to 17	26	38.5	19	42.1	7	28.6	60	36.7	46	37.0	14	35.7
18 to 25	13	46.2	10	60.0	3	–	24	29.2	12	25.0	12	33.3
26 to 49	15	13.3	11	18.2	4	–	6	16.7	3	33.3	3	–
50 or Older	7	14.3	7	14.3	0	–	0	–	0	–	0	–
Total	61	31.1	47	32.0	14	14.3	90	33.3	61	34.4	29	31.0

Age of Respondent	Sedatives					
	Total Triggered	Percent Converted	Triggered from "Don't Know"	Percent Converted	Triggered from "Refused"	Percent Converted
12 to 17	9	44.4	7	57.1	2	–
18 to 25	9	77.8	6	83.3	3	66.7
26 to 49	2	50.0	2	50.0	0	–
50 or Older	1	–	1	–	0	–
Total	21	57.1	16	62.5	5	40.0

Table 8.9 Responses to the 30-Day Frequency Follow-Up Items
(Summarized Across All Drug Types)

Response Given to Follow-Up	Overall		"Don't Know"		"Refused"	
	Percent	*N*	Percent	*N*	Percent	*N*
1 of the 6 Frequency Categories	79.7	802	85.3	704	54.2	98
Refused	8.4	84	1.1	9	41.4	75
Don't Know	11.9	120	13.6	112	4.4	8
Total	100.0	1,006	100.0	825	100.0	181

When comparing the success of this follow-up with the other two types of follow-ups (lifetime and recency), it is important to note that this one is unique in that it asks respondents to make their best estimate, then provides a set of answer categories rather than having the respondents recall a specific number of days. Thus, this is the only follow-up in which the follow-up item actually simplifies the respondent's reporting task. With the lifetime and recency follow-ups, the respondent is essentially asked to "try harder" or "give us a guess" with no additional assistance. Task simplification appears to play a significant role in the success of the follow-up methodology.

Table 8.10 shows that although the 30-day follow-up was effective for respondents of all ages, it too was most effective for the younger respondents. Respondents aged 12 to 17 provided a substantive response 82 percent of the time, and respondents aged 18 to 25 provided a substantive response 81 percent of the time. This compares with rates of 75 and 70 percent for 26 to 49 year olds and respondents aged 50 or older, respectively. The differential conversion rate by age group is less pronounced than for the lifetime use follow-up, but the same pattern appears.

Table 8.10 Respondents Converted by the Follow-Up Items for 30-Day Frequency, by Age
(Summarized Across All Drug Types)

Respondent's Age Group	Number of Follow-Up Respondents	Respondents Converted	
		Percent	*N*
12 to 17	477	81.6	389
18 to 25	330	80.9	267
26 to 49	140	75.0	105
50 or Older	59	69.5	41

The distribution of responses to the 30-day frequency follow-up does not indicate systematic variation in results by drug type. ***Table 8.11*** provides results for cigarettes, alcohol, marijuana, and cocaine. The other drugs to which the follow-up applied (crack, heroin, hallucinogens, and inhalants) resulted in too few observations to interpret the results meaningfully. For almost every drug listed in ***Table 8.11***, the rate of conversion to a substantive response was on a par with the overall average and the average according to age group. Results varied as the sample sizes became smaller for marijuana and cocaine. For example, the overall conversion rate for cocaine was 60 percent compared with 85, 78, and 75 percent for cigarettes, alcohol, and marijuana, respectively. However, the sample size was much smaller for cocaine, and the results should be interpreted with care. In summary, the 30-day frequency follow-up was judged a success, as it converted a high percentage of both "Don't knows" and refusals.

Table 8.11 Drug-Specific Results for the 30-Day Frequency Follow-Up, by Age of Respondent

	Cigarettes		Alcohol		Marijuana		Cocaine (Including Crack)	
Response	**Percent**	***N***	**Percent**	***N***	**Percent**	***N***	**Percent**	***N***
Total	100.0	384	100.0	334	100.0	111	100.0	10
1 of 6 Frequency Periods	84.6	325	78.1	261	74.8	83	60.0	6
Refused	3.7	14	10.2	34	16.2	18	30.0	3
Don't Know	11.7	45	11.7	39	9.0	10	10.0	1
12 to 17	100.0	191	100.0	122	100.0	65	100.0	4
1 of 6 Frequency Periods	85.9	164	77.1	94	81.5	53	75.0	3
Refused	1.1	2	11.5	14	10.8	7	25.0	1
Don't Know	13.1	25	11.5	14	7.7	5	0.0	0
18 to 25	100.0	140	100.0	103	100.0	32	100.0	3
1 of 6 Frequency Periods	85.7	120	84.5	87	65.6	21	33.3	1
Refused	5.0	7	5.8	6	25.0	8	33.3	1
Don't Know	9.3	13	9.7	10	9.4	3	33.3	1
26 to 49	100.0	42	100.0	68	100.0	11	100.0	3
1 of 6 Frequency Periods	78.6	33	75.0	51	72.7	8	66.7	2
Refused	7.1	3	14.7	10	27.3	3	33.3	1
Don't Know	14.3	6	10.3	7	0.0	0	0.0	0
50 or Older	100.0	11	100.0	41	100.0	3	0.0	0
1 of 6 Frequency Periods	72.7	8	70.7	29	33.3	1	0.0	0
Refused	18.2	2	9.8	4	0.0	0	0.0	0
Don't Know	9.1	1	19.5	8	66.7	2	0.0	0

Summary of Descriptive Analyses. The results presented thus far show that item nonresponse for this subset of items in the 2000 instrument was quite small; the overwhelming majority of respondents triggered none of the follow-up items. Among respondents who did trigger follow-up items, conversion to a substantive response varied by the type of follow-up question. For the lifetime refusal follow-up, less than half of the respondents provided a substantive response. Youth respondents were more likely than older respondents to convert their responses to a substantive answer. A slightly higher percentage of recency follow-up respondents converted to a substantive response. Although youth respondents were still more likely to be converted, the differences were much smaller than for the lifetime use follow-up items. The 30-day frequency follow-up items were much more successful, converting about 80 percent to a substantive response. This item is unique among the three in that it alters the response task for the respondent by facilitating recall. It provides categories from which the respondent can select rather than simply asks the respondent to answer the same question. The fact that, in terms of conversion rates, the most successful follow-up questions were those that simplified the reporting task suggests that the follow-up methodology may be most effective for converting "Don't know" responses rather than refusals. This is consistent with the finding from both the recency and lifetime follow-up items that conversion rates for initial refusals were fairly low.

Multivariate Models for Triggering Follow-Up Questions

To determine what other characteristics of respondents tend to be associated with triggering follow-up questions, several multivariate models were developed. Logistic regression was used

to determine the likelihood of triggering a follow-up (e.g., answering "Don't know" or "Refused" to a critical lifetime, recency, or frequency drug question). The lifetime follow-up model was run on all cases, while the recency and frequency follow-up models were run on only those subsets of respondents who reported lifetime use (in the case of the recency follow-up item) or who reported use in the past 30 days (in the case of the frequency follow-up). ***Table 8.12*** describes the predictor variables included in these models.

Based on the results of the descriptive analyses, age was expected to be a fairly consistent predictor of answering "Don't know" or "Refused" to critical drug items. Men were expected to trigger follow-ups more often than women. Furthermore, respondents who had been less willing to participate in the study (as measured by whether a sample person refusal was ever recorded in the record of calls) were expected to be more likely to give a "Don't know" or "Refused" answer to a critical question during the interview. Region of residence and population density also were included in the models with the expectation that respondents residing in areas of the country with higher unit nonresponse (e.g., the Northeast) and more heavily populated areas would be more likely to trigger follow-up questions. Comprehension, cooperativeness, and privacy were all expected to be negatively associated with triggering a follow-up question. In contrast, sharing answers with the interviewer was expected to be positively associated as that would be viewed as an indication of lesser privacy concerns on the part of the respondent.

Table 8.12 Predictor Variables Included in the Logistic Models

Variable Description	Categories
Age of Respondent	12 to 17 18 to 25 26 to 49 50 or Older*
Gender of Respondent	Male Female*
Race/Ethnicity of Respondent	Hispanic Non-Hispanic Black Non-Hispanic Non-Black*
Education of Respondent	12 to 17 Years old Less Than High School High School Graduate More Than High School*
Household Income	More Than $20,000 $20,000 or Less*
Region of Residence	Northeast Midwest South West*
Population Density	MSA ≥1,000,000 MSA < 1,000,000 Non-MSA*
Pending Refusal (was a status code of "pending refusal" ever assigned for this respondent?)	Yes No*
Respondent Comprehension (interviewer's assessment of respondent comprehension during the interview)	No Difficulty At Least Some Difficulty A Lot of Difficulty*
Respondent Cooperation (interviewer's assessment of respondent cooperation during the interview)	Cooperative Not Cooperative*
Privacy (interviewer's assessment of the privacy of the interview setting)	Completely Private Not Private*
Sharing Answers (interviewer's report of frequency with which the respondent shared his or her answers with the interviewer)	None of the Time At Least Some of the Time All of the Time*

*Denotes the reference category.
MSA = metropolitan statistical area.

Several different logistic models were developed. The predictor variables were used to explain the type of follow-up question triggered, as well as the type of drug for which the follow-up was triggered. That is, the models were used to gauge the independent effects of these

variables on triggering any type of follow-up response for any drug, the lifetime follow-up response for any drug, the recency follow-up response for any drug, and the frequency follow-up response for any drug. In addition, models for triggering the recency and frequency items were analyzed separately for the individual drugs (cigarettes, alcohol, and marijuana). A drug-specific model for triggering the lifetime item was run only for marijuana–too few respondents triggered the lifetime item for each of the other drug types.

Table 8.13 presents a matrix of characteristics found to be statistically significant in the models that were created to explain the triggering of the follow-up items. Each of the models presented in the table is described in detail below.

Triggering Any Type of Follow-Up Item. First, the most general model was tested to determine the likelihood of triggering any follow- up item across any of the three follow-up question types (lifetime, recency, and frequency) over all drug types. Five predictor variables yielded statistically significant results at the 0.05 level. These were age group, gender, education, region of residence, and the interviewer's perception of the respondent's level of cooperation. This general model suggests that 12 to 17 year olds were 1.7 times more likely to trigger any type of item nonresponse follow-up for any drug than were respondents aged 50 or older. Male respondents also were more likely to trigger the follow-up questions—at a level 1.5 times that of females. Respondents with only a high school degree were 1.3 times more likely to trigger the follow-ups than respondents with more than a high school education. Respondents in the South were 1.3 times more likely to trigger a follow-up than respondents in the West. Finally, as expected, respondents who were categorized as "cooperative" by their interviewer were about two thirds *less* likely to trigger an item nonresponse follow-up question for any type of drug.

Triggering the Lifetime Refusal Follow-Up. The general model for lifetime use also identified statistically significant (at the 0.05 level) findings for male respondents (twice as likely as female respondents to trigger a lifetime use follow-up item) and cooperative respondents (80 percent less likely than noncooperative respondents to trigger a lifetime follow-up item). Unlike the general model, no statistically significant results were found for age group, education level, or region. Due to the small numbers of respondents triggering a follow-up for lifetime use of most substances, the only drug analyzed separately was marijuana. For marijuana, the same two variables yielded statistically significant results. Male respondents were again about twice as likely to trigger the follow-up as female respondents. Respondents who were coded as "cooperative" also were about 80 percent less likely than noncooperative respondents to trigger the lifetime refusal follow-up for marijuana.

Triggering the Recency Follow-Up. For the most part, the same variables yielded statistically significant results in the model for triggering the recency follow-up for any drug as they did in triggering the lifetime follow-up for any drug. In the age category, 12 to 17 year olds were 3.8 times more likely to trigger a recency follow-up than were respondents aged 50 or older. Male respondents were again 1.5 times more likely to trigger a follow-up than female respondents. With respect to education, respondents with no more than a high school degree (100 percent) were more likely to trigger the recency follow-up than respondents with more than a high school degree. Finally, respondents in the Midwest were about a third less likely to trigger a recency follow-up than respondents in the West.

Table 8.13 Significant Predictors for Triggering Follow-Up Questions

Drug	Significant Characteristic	Odds Ratio (95 Percent Confidence Interval)		*P* Value
Overall (All Types of Follow-Ups)				
All Drugs	12 to 17 Years Old (+)	1.68	(1.21,2.35)	0.0022
	Male (+)	1.50	(1.23,1.83)	<0.0001
	Cooperative (-)	0.37	(0.21,0.68)	0.0012
	High School Only (+)	1.34	(1.05,1.70)	0.0192
	South (+)	1.31	(1.00,1.71)	0.0478
Cigarettes	26 to 49 Years Old (-)	0.52	(0.31,0.88)	0.0156
	Male (+)	1.41	(1.00,1.99)	0.0478
	South (+)	1.89	(1.15,3.10)	0.0125
Alcohol	26 to 49 Years Old (-)	0.50	(0.30,0.82)	0.0065
	Midwest (-)	0.54	(0.32,0.93)	0.0262
	Cooperative (-)	0.21	(0.09,0.48)	0.0002
Marijuana	Male (+)	2.03	(1.19,3.46)	0.0089
	Cooperative (-)	0.17	(0.05,0.50)	0.0015
Lifetime Use Follow-Up				
All Drugs	Male (+)	2.08	(1.29,3.36)	0.0025
	Cooperative (-)	0.19	(0.07,0.53)	0.0014
Marijuana	Male (+)	2.12	(1.14,3.93)	0.0178
	Cooperative (-)	0.17	(0.05,0.60)	0.0061
Recency of Use Follow-Up				
All Drugs	12 to 17 Years Old (+)	3.75	(2.30,6.13)	<0.0001
	Male (+)	1.46	(1.13,1.89)	0.0041
	High School Only (+)	2.02	(1.43,2.86)	<0.0001
	Midwest (-)	0.65	(0.43,0.96)	0.0316
Cigarettes	26 to 49 Years Old (-)	0.30	(0.13,0.67)	0.0036
	Cooperative (+)	8.34	(1.03,67.37)	0.0466
Alcohol	12 to 17 Years Old (+)	5.13	(1.37,19.17)	0.0151
	High School Only (+)	3.40	(1.14,10.16)	0.0281
	Income >$20K (-)	0.30	(0.11,0.81)	0.0175
	No Difficulty Understanding Interview (-)	0.29	(0.14,0.60)	0.0008
Marijuana	Male (+)	1.86	(1.04,3.33)	0.0352
	Black (+)	2.32	(1.06,5.11)	0.0359
	Less Than High School (+)	2.24	(1.03,4.36)	0.0412
	Midwest (-)	0.42	(0.20,0.92)	0.0299
	High Population Density (-)	0.50	(0.26,0.96)	0.0366
30-Day Frequency of Use Follow-Up				
All Drugs	12 to 17 Years Old (+)	3.72	(2.26,6.12)	<0.0001
	26 to 49 Years Old (-)	0.55	(0.35,0.87)	0.0106
	Black (+)	2.12	(1.47,3.05)	<0.0001
	Cooperative (-)	0.23	(0.11,0.51)	0.0003
Cigarettes	12 to 17 Years Old (+)	5.98	(2.07,17.27)	0.0010
	Black (+)	1.97	(1.18,3.30)	0.0010
	High Population Density (-)	0.64	(0.42,0.99)	0.0463
	Moderate Population Density (-)	0.58	(0.37,0.91)	0.0173
Alcohol	12 to 17 Years Old (+)	2.33	(1.22,4.44)	0.0101
	18 to 25 Years Old (-)	0.57	(0.33,0.96)	0.0360
	26 to 49 Years Old (-)	0.36	(0.21,0.64)	0.0005
	Black (+)	2.47	(1.46,4.16)	0.0008
	Cooperative (-)	0.18	(0.07,0.44)	0.0002
Marijuana	Black (+)	2.15	(1.13,4.10)	0.0200
	Cooperative (-)	0.12	(0.02,0.81)	0.0294

Note: Table provides respondent characteristics, statistically significant at 0.05 level, by drug type and question type (+ more likely to trigger follow-up; - less likely to trigger follow-up).

[1] Sample sizes vary by category of follow-up: The overall and lifetime categories include all respondents, recency includes only respondents who reported lifetime use, and 30-day frequency includes only those respondents who reported past 30-day use.

Unlike the lifetime model, the recency model produced different results by drug type, however. For cigarettes, 26 to 49 year olds were only 30 percent as likely as respondents aged 50 or older to trigger a nonresponse follow-up question. Oddly, cooperative respondents were over 8 times *more* likely to trigger the recency follow-up than noncooperative respondents for the cigarette question. For alcohol, respondents aged 12 to 17 were 5 times more likely to trigger the recency follow-up than respondents aged 50 or older. Respondents with only a high school degree were 3.4 times more likely to trigger the follow-up than respondents with more than a high school degree. Respondents with an income greater than $20,000 were only 30 percent as likely as those with an income of less than $20,000 to trigger a recency follow-up. Finally, respondents who were coded as having "no difficulty" completing the interview were only about 30 percent as likely as respondents who had at least some difficulty in triggering a recency follow-up. For marijuana, there were five statistically significant predictor variables. Males were nearly twice as likely as females to trigger the follow-up, black respondents were 2.3 times more likely than whites to trigger the follow-up, respondents with less than a high school degree were 2.2 times more likely to trigger the follow-up, Midwesterners were only 42 percent as likely as respondents in the western United States to trigger a recency follow-up, and respondents in metropolitan statistical areas (MSAs) greater than or equal to 1,000,000 were 50 percent less likely to trigger it than respondents in non-MSAs.

Statistically significant predictor variables were varied for the recency item. The data showed no consistent pattern across types of drugs. One explanation for this could be the fact that the recency item is triggered through both "Don't know" and "Refused" responses. Thus, a wider range of factors is likely to lead to each type of response, including varying levels of comfort reporting recency of use of various drugs and legitimately not knowing the last time a drug was used. A larger sample size would enable the generation of separate models for the "Don't know" and "Refused" responses that would further clarify these findings.

Triggering the 30-Day Frequency Follow-Up. Over all drugs, the frequency follow-up was 78 percent more likely to be triggered among 12 to 17 year olds than it was among respondents aged 50 or older, 58 percent more likely to be triggered among black than white respondents, and 73 percent less likely to be triggered among respondents who were coded as "cooperative." Age, race/ethnicity, and population density were statistically significant variables for the cigarette frequency follow-up item. Youth respondents were 6 times more likely than respondents aged 50 or older to trigger the follow-up. Black respondents were 2 times more likely than white respondents to trigger the follow-up. Respondents living in MSAs with a population of 1,000,000 or more were 36 percent less likely, and respondents in MSAs with fewer than 1,000,000 were 42 percent less likely than respondents in non-MSAs to trigger a follow-up. For alcohol, all age categories were statistically significant at the 0.05 level. Youths were 2.3 times more likely to trigger a follow-up than respondents aged 50 or older, but 18 to 25 year olds (43 percent less likely) and 26 to 49 year olds (64 percent less likely) were both less likely than those aged 50 or older to trigger the follow-up. Black respondents were 2.5 times more likely to trigger a follow-up than white respondents. Furthermore, respondents coded as "cooperative" were 82 percent less likely to trigger a follow-up than respondents coded as "noncooperative." The marijuana frequency follow up generated only two variables that were statistically significant. Black respondents were twice as likely as white respondents to result in the frequency follow-up. Cooperative respondents were 88 percent less likely to trigger an item nonresponse follow-up for frequency.

Variables affecting the 30-day frequency follow-up appeared to be more consistent. Black respondents, youth respondents, and cooperative respondents were more likely to trigger the 30-day frequency follow-up in most of these models. The consistency of these findings may be a result of the likely reason that respondents trigger this type of follow-up. In this case, triggering the follow-up is less an issue of sensitivity and more an issue of recall difficulties.

Triggering Follow-Ups, by Drug Type. To determine whether any particular drug exhibited different relationships between respondent characteristics and the likelihood of triggering a follow-up, separate models were run for cigarettes, alcohol, and marijuana across all three types of follow-up questions (see ***Table 8.13***).

For cigarettes, 26 to 49 year olds were only half as likely as respondents aged 50 or older to trigger any type of item nonresponse follow-up. Male respondents were about 1.4 times more likely than female respondents to trigger any type of follow-up. Also for cigarettes, respondents in the South were 89 percent more likely than respondents in the West to trigger any type of follow-up question. For alcohol, 26 to 49 year olds were again only half as likely as respondents aged 50 or older to trigger a follow-up. Respondents in the Midwest also were only about half as likely (54 percent) as respondents in the West to trigger a follow-up. Furthermore, respondents coded as "cooperative" were 79 percent less likely than respondents coded as "noncooperative" to trigger a follow-up. For marijuana, two variables were statistically significant across all types of follow-up questions. Male respondents were twice as likely as female respondents to trigger any type of follow-up. Respondents coded as "cooperative" were 17 percent as likely as respondents coded "noncooperative" to trigger any type of follow-up. From these results, it is difficult to determine any difference in response by drug type.

Examining the effects of respondent characteristics on triggering follow-ups by drug type, and across the three different types of follow-ups, the most consistent predictors were the respondent's gender and level of cooperation with the interviewer. Male respondents were more likely to trigger follow-ups across all drug and question types. This was expected, however, because males are more likely to be drug users. As noted earlier, use was controlled by running the recency models only for respondents who reported lifetime use and the frequency models only for respondents who reported 30-day use. As such, there may be an additional propensity among male respondents to avoid disclosing sensitive information. Cooperative respondents, as expected, were far less likely to trigger follow-ups than noncooperative respondents. The fact that this variable was found to be a statistically significant predictor provides some face validity to the use of the interviewers' observations to measure cooperation.

Multivariate Models for Converting to a Substantive Response

In addition to modeling predictors of triggering the item nonresponse follow-ups, factors associated with resolution of the follow-ups were examined. Using the same predictor variables, the likelihood of converting an initial "Don't know" or "Refused" to one of the substantive response categories for that question was modeled. For these models, analyses were restricted to only one type of question and one drug type at a time. This helped to avoid the problem of undercounting cases in situations where a respondent triggered more than one type of follow-up question in more than one drug category.

Table 8.14 presents a matrix of characteristics found to be statistically significant in the models that were created to explain the resolution of the follow-up items. Each of the models presented in the table is described in detail below.

Resolving the Lifetime Refusal Follow-Up. There were only a sufficient number of cases to run the lifetime refusal follow-up model for marijuana. Three variables did achieve statistical significance at the 0.05 level. Respondents aged 18 to 25 were 21 times more likely than respondents aged 50 or older to convert to a substantive answer, and respondents aged 26 to 49 were 9 times more likely than respondents aged 50 or older to convert to a substantive answer after being exposed to the follow-up. Respondents who had ever refused the interview were only 10 percent as likely to convert to a valid answer as respondents who had never refused to participate in the interview. Finally, respondents from the Northeast (4 percent as likely) and the Midwest (16 percent as likely) were both far less likely to convert to a valid answer than respondents from the West.

Resolving the Recency Follow-Ups. Respondents who answered "Don't know" or "Refused" to the recency item for cigarettes, alcohol, or marijuana were modeled separately by drug. For the cigarette model, age, gender, education, and whether or not the respondent had let the field interviewer (FI) know his or her answers were all statistically significant at the 0.05 level. Youth respondents (12 to 17 years old) were 18 times more likely to resolve their answers than respondents aged 50 or older. Males were over 5 times more likely than females to resolve their answers. Respondents with only a high school level of education were 11.4 times more likely than respondents with an education beyond high school to resolve their answers. Finally, respondents who never let the FI know their answers were almost 10 times more likely to resolve to a substantive response.

For the alcohol model, males were again nearly 5 times more likely than females to resolve their answers, and respondents with a high school education were 8 times more likely to resolve their answers than respondents with an education beyond high school. Midwesterners were 11 times more likely than Westerners to resolve their answers, and interestingly, respondents for whom the interview was completely private were 73 percent *less likely* to resolve their answers. With respect to marijuana, there were five statistically significant variables. Black respondents were 83 percent less likely to resolve their answers than whites, respondents with incomes over $20,000 were 84 percent less likely to resolve their answers than respondents with incomes of less than $20,000, and Northeasterners were 4 times more likely to resolve to a substantive answer than were Westerners. Respondents for whom the interview was completely private were almost 5 times more likely to resolve their answers than it was for those whose interview was not private. Finally, respondents for whom the FI never knew the answers were only 17 percent as likely to resolve to a substantive answer.

Resolving the 30-Day Frequency Follow-Ups. The frequency items were modeled separately for cigarettes and alcohol. For cigarettes, respondents living in MSAs with fewer than 1,000,000 people were 4.2 times more likely than respondents in non-MSAs to resolve the frequency item for cigarettes. Respondents who had no difficulty understanding the interview were 3 times more likely than respondents who had difficulty resolving their answers. Only two statistically significant variables were found for alcohol as well. Respondents in the Northeast were only 21 percent as likely as respondents in the West to resolve their answers, and respondents in MSAs

Table 8.14 Predicting Resolution of Follow-Up Questions

Drug	Significant Characteristic	Odds Ratio (95 Percent Confidence Interval)		*P* Value
Lifetime Follow-Up[1]				
Cigarettes	N/A	N/A	N/A	N/A
Alcohol	N/A	N/A	N/A	N/A
Marijuana	18 to 25 Years Old (+)	21.18	(2.70,166.09)	0.0037
	26 to 49 Years Old (+)	9.09	(1.57,52.52)	0.0137
	Ever Refused (-)	0.10	(0.01,0.81)	0.0310
	Northeast (-)	0.04	(0.01,0.30)	0.0018
	Midwest (-)	0.16	(0.03,0.94)	0.0431
Recency Follow-Up				
Cigarettes	12 to 17 Years Old (+)	17.96	(1.35,238.21)	0.0286
	Male (+)	5.27	(1.37,20.24)	0.0155
	High School Only (+)	11.41	(1.49,87.37)	0.0192
	Field Interviewer Never Knew Answers (+)	9.59	(1.12,82.42)	0.0395
Alcohol	Male (+)	4.75	(1.64,13.72)	0.0041
	High School Only (+)	7.95	(1.22,51.80)	0.0302
	Midwest (+)	10.88	(1.28,92.36)	0.0288
	Completely Private Interview (-)	0.27	(0.08,0.88)	0.0301
Marijuana	Black (-)	0.17	(0.04,0.75)	0.0191
	Income > $20K (-)	0.16	(0.04,0.62)	0.0084
	Northeast (+)	4.22	(1.07,16.57)	0.0393
	Completely Private Interview (+)	4.75	(1.01,22.36)	0.0489
	Field Interviewer Never Knew Answers (-)	0.17	(0.03,0.96)	0.0446
30-Day Frequency Follow-Up				
Cigarettes	Moderate Population Density (+)	4.22	(1.37,13.00)	0.0122
	No Difficulty Understanding Interview (+)	3.05	(1.01,9.22)	0.0486
Alcohol	Northeast (-)	0.21	(0.05,0.89)	0.0342
	Moderate Population Density (+)	5.13	(1.55,16.95)	0.0074
Marijuana	N/A	N/A	N/A	N/A

N/A = not applicable.

Note: Table provides respondent characteristics, statistically significant at 0.05 level, by drug type and question type (+ more likely to resolve follow-up; - less likely to resolve follow-up).

[1] Sample sizes vary by category of follow-up: The lifetime follow-up includes all respondents, recency includes only respondents who reported lifetime use, and 30-day frequency includes only those respondents who reported past 30-day use.

with fewer than 1,000,000 people were 5 times more likely than respondents outside MSAs to resolve to a substantive answer.

Overall, even less of a pattern appeared across drug and question types with the resolution models than with the triggering models. Fewer of the same variables (such as age, race/ethnicity, and region) were statistically significant at the 0.05 level. Other variables emerged for the follow-up items, but no clear pattern was apparent.

Effect of the Follow-Up Methodology on Lifetime Prevalence Estimates

The main reason for developing the follow-up methodology was to reduce nonresponse to critical items. This is especially important to the extent that use patterns differ between respondents who provide a substantive response to the original item and those who provide a substantive response only when prompted with the follow-up item. The descriptive results presented earlier show that among respondents who were routed to the lifetime follow-up item and who provided a substantive response, a larger percentage reported use than did not (65 vs. 35 percent). The final analyses of this chapter assess the impact of the follow-up responders on the prevalence estimates for lifetime use generated from the 2000 data. It was hypothesized that respondents who provided "Refused" responses to the lifetime use items were doing so as a means to avoid disclosing a sensitive behavior. Thus, follow-up responders would be expected to report higher rates of drug use than the respondents who answered the original question. ***Tables 8.15*** and ***8.16*** provide these data for each drug that included a lifetime follow-up item. The data in ***Table 8.15*** are unweighted, and those in ***Table 8.16*** are weighted. Data are shown only for the full sample as cell sizes become too small to interpret the age subgroups.

Table 8.15 Unweighted Prevalence Estimates for Lifetime Use

	Prevalence Estimate				Sample Size			
Substance	Gate Only	Follow-Up Only	Gate and Follow-Up	Imputed Revised[1]	Gate Only	Follow-Up Only	Gate and Follow-Up	Imputed Revised[1]
Cigarettes	56.99	48.28	56.99	56.99	71,735	29	71,764	71,764
Alcohol	70.06	51.61	70.05	70.05	71,702	31	71,733	
Marijuana	33.32	81.61	33.38	33.41	71,553	87	71,640	
Cocaine	8.48	65.00	8.50	8.51	71,680	20	71,700	
Crack[2]	2.06	100.00	2.07	2.07	71,689	3	71,692	
Heroin	0.93	50.00	0.93	0.94	71,721	2	71,723	
LSD	8.64	62.50	8.65	8.65	71,677	8	71,685	
PCP	1.91	16.67	1.91	1.91	71,671	6	71,677	

[1]Includes edits and imputations.

[2]Entry into the crack module is contingent on at least lifetime use of cocaine. Respondents who reported never having used cocaine were logically defined to have never used crack.

Looking first at ***Table 8.15***, the data show that prevalence rates were actually lower among follow-up responders for cigarettes and alcohol. However, the rates were higher among the follow-up responders for the remaining drug types. Caution should be used in interpreting the data for heroin, LSD, and PCP as sample sizes are quite small. The data for marijuana and cocaine are based on somewhat larger sample sizes and suggest that for the more "serious" drugs, the follow-up methodology may be contributing to a more accurate estimate of lifetime prevalence (based on the assumption that a higher prevalence rate is more accurate). The results in ***Table 8.16*** are similar, showing that the results are not simply an artifact of the weighting procedures.

Table 8.16 Weighted Prevalence Estimates for Lifetime Use

Substance	Prevalence Estimate				Sample Size			
	Gate Only	Follow-Up Only	Gate and Follow-Up	Imputed Revised[1]	Gate Only	Follow-Up Only	Gate and Follow-Up	Imputed Revised[1]
Cigarettes	66.46	40.73	66.45	66.45	71,735	29	71,764	71,764
Alcohol	80.97	53.24	80.96	80.97	71,702	31	71,733	
Marijuana	34.06	85.51	34.13	34.18	71,553	87	71,640	
Cocaine	11.12	55.86	11.13	11.15	71,680	20	71,700	
Crack[2]	2.36	100.00	2.36	2.38	71,689	3	71,692	
Heroin	1.24	90.54	1.24	1.24	71,721	2	71,723	
LSD	8.78	77.90	8.80	8.80	71,677	8	71,685	
PCP	2.60	21.26	2.60	2.60	71,671	6	71,677	

[1]Includes edits and imputations.

[2]Entry into the crack module is contingent on at least lifetime use of cocaine. Respondents who reported never having used cocaine were logically defined to have never used crack.

Although the lifetime prevalence rates for marijuana and cocaine were much higher among the follow-up responders, they contributed virtually nothing to the overall estimates derived from the 2000 survey data. Given the small number of respondents who actually triggered the follow-up items and the large sample size associated with the 2000 survey, it is impossible for the estimates to change much if at all. Thus, although there is some evidence to suggest that the follow-up responders may be more likely to be drug users, their inclusion in the final estimates did not change the estimate from that based on only those respondents who answered the original item.

Conclusions

Perhaps the most significant finding from these analyses is that item nonresponse to these critical items was quite low. For the most part, respondents were willing to answer these questions and did not require additional prompting to do so. As a result of the low item nonresponse rates, the data presented here must be interpreted with care. The results of these analyses suggest that younger respondents were more likely to trigger the follow-ups and to provide substantive responses to the follow-ups. In addition, the follow-up methodology was more successful in converting respondents who triggered the follow-up through a "Don't know" response than through a "Refused" response. The methodology also was more successful when combined with a revised question that reduced respondent recall burden as was done with the 30-day frequency follow-up for "Don't know." The largest percentage of follow-up responders provided a substantive response to the 30-day frequency follow-up when the question was simplified by providing response categories in place of the open-ended response field. There also was some evidence to suggest that drug use may be more prevalent among the follow-up responders although small sample sizes precluded a thorough examination of this result.

Taken together, these results suggest that the follow-up methodology is a useful strategy for reducing item nonresponse—particularly when the nonresponse is due to "Don't know" responses. Additional thought should be given to whether improvements can be made to the

refusal follow ups to increase the number of respondents who convert to a substantive response. Focus groups could be useful in identifying other reasons (beyond the fear of disclosure and questions about the importance of the data) that could cause respondents to refuse these critical items. The results of such focus groups could be used to develop more appropriately worded follow-ups that might be more effective in persuading respondents to provide substantive responses.

Future users of this methodology would be wise to include follow-up items for questions that will be used to define whether a case will be defined as complete, and as such, included in the final analysis file for the study. Given the high cost of conducting face-to-face interviews, the opportunity to avoid discarding even a small number of completed cases by using this follow-up methodology could be substantial.

References

Alwin, D. F. (1977). Making errors in surveys: An overview. *Sociological Methods and Research, 6*, 131-150.

Bailar, B. A., Bailey, L., & Stevens, J. (1977). Measures of interviewer bias and variance. *Journal of Marketing Research, 14*, 337-343.

Craig, C. S., & McCann, J. M. (1978). Item nonresponse in mail surveys: Extent and correlates. *Journal of Marketing Research, 15*, 285-289.

de Leeuw, E. D. (1992). *Data quality in mail, telephone, and face to face surveys*. Amsterdam, The Netherlands: TT-Publikaties.

de Leeuw, E. D. (1999). Item non-response: Prevention is better than cure. *Survey Methods Newsletter, 19*(2), 4-8.

DeMaio, T. (1984). Social desirability and survey measurement: A review. In C. F. Turner & E. Martin (Eds.), *Surveying subjective phenomena* (pp. 257-282, Volume II). New York: Russell Sage Foundation.

Duffer, A., Lessler, J., Weeks, M., & Mosher, W. (1996). Impact of incentives and interviewing modes: Results from the National Survey of Family Growth Cycle V pretest. In R. Warnecke (Ed.), *Health survey research methods* (pp. 147-152). Hyattsville, MD: National Center for Health Statistics.

Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.

Jenkins, C. R., & Dillman, D. A. (1997). Towards a theory of self-administered questionnaire design. In L. Lyberg , P. Biemer, M. Collins, L. Decker, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 165-196). New York: Wiley-Interscience.

Lessler, J. T., & Kalsbeek, W. D. (1992). *Nonsampling error in surveys*. New York: Wiley.

Office of Applied Studies. (2000). *Summary of findings from the 1999 National Household Survey on Drug Abuse* (DHHS Publication No. SMA 00-3466, NHSDA Series H-12). Rockville, MD: Substance Abuse and Mental Health Services Administration.

Rogers, S. M., Miller, H. G., Forsyth, B. H., Smith, T. K., & Turner, C. F. (1996). Audio-CASI: The impact of operational characteristics on data quality. In *Proceedings of the American Statistical Association, Survey Research Methods Section, Volume II* (pp. 1042-1047). Alexandria, VA: American Statistical Association.

Sudman, S., Bradburn, N. M., Blair, E., & Stocking, C. (1977). Modest expectation: The effects of interviewers' prior expectations on responses. *Sociological Methods and Research, 6*, 171-182.

Turner, C., Ku, L., Sonenstein, F., & Pleck, J. (1996). Impact of ACASI on reporting of male-male sexual contacts: Preliminary results from the 1995 National Survey of Adolescent Males. In R. Warnecke (Ed.), *Health survey research methods* (pp. 171-176). Hyattsville, MD: National Center for Health Statistics.

Turner, C. F., Lessler, J. T., & Gfroerer, J. C. (Eds.). (1992). *Survey measurement of drug use: Methodological studies* (DHHS Publication No. ADM 92-1929). Rockville, MD: National Institute on Drug Abuse.

9. A Test of the Item Count Methodology for Estimating Cocaine Use Prevalence

Paul P. Biemer and B. Kathleen Jordan
RTI International

Michael Hubbard
University of North Carolina at Chapel Hill

Douglas Wright
Substance Abuse and Mental Health Services Administration

Introduction

The Substance Abuse and Mental Health Services Administration (SAMHSA) has long sought ways to improve the accuracy of the prevalence estimates provided by the National Survey on Drug Use and Health (NSDUH).[1] One method of data collection that shows some promise for improving reporting accuracy is the "item count method." This technique provides respondents with an enhanced perception of anonymity when reporting a sensitive behavior, such as drug use. This is accomplished by including the sensitive behavior of interest in a list of other relatively nonstigmatizing behaviors. The respondent reports the number of items in the list in which he or she has engaged. Only the number of behaviors is reported, not which specific behaviors apply. Because the report does not allow anyone accessing the data to know which specific behaviors are true for a respondent, there is no way to determine whether a respondent has admitted to the sensitive behavior. However, because the average number of nonsensitive behaviors can be estimated for the population, one can estimate the rate of the sensitive behavior for the population by the difference between the average number of behaviors reported for the population including and excluding the stigmatized behavior.

To test the efficacy of the item count (IC) methodology for estimating drug use prevalence, the method was implemented in the 2001 survey. This chapter describes the research conducted in 2000 and 2001 to (1) develop an IC module for past year cocaine use, (2) evaluate and refine the module using cognitive laboratory methods, (3) develop the prevalence estimators of past year cocaine use for the survey design, and (4) make final recommendations on the viability of using the IC method to estimate drug use prevalence. As part of (4), IC estimates of past year cocaine use based upon this implementation are presented, and the validity of the estimates is discussed.

Past experience with the IC methodology (e.g., Droitcour et al., 1991) has identified two major problems with this method: task difficulty and the selection of the innocuous items for the

[1] Prior to 2002, the survey was called the National Household Survey on Drug Abuse (NHSDA). In this report, it is referred to as NSDUH, regardless of the survey year.

list. Regarding task difficulty, respondents have had difficulty, and have often made errors, computing the number of activities in which they have engaged. There appeared to be several reasons for this. First, some respondents became confused by the question and gave the item number of a particular activity that applied rather than counting up all the activities that applied. Second, the process of deciding which items apply and keeping a running tally of these as the list is read has proven difficult for some respondents. Respondents may have made mistakes and entered an erroneous count or may have missed some items that applied as they mentally calculated a running total.

With respect to selection of the innocuous items, the "nonstigmatizing" items in the list have often been common activities or conditions that were unrelated to the subject matter of the questionnaire. As examples, the items have asked about food the respondent may have eaten, locations visited, and automobiles purchased. The introduction of a stigmatizing behavior among these innocuous behaviors could make the respondent suspicious that there was some trick involved and that the investigators would in some way be able to determine whether the respondent had engaged in the sensitive behavior. Consequently, some respondents may have deliberately misreported the number of behaviors that applied.

Probably as a result of these problems, it appeared that the accuracy of prevalence rates developed using this methodology was no better, and possibly worse, than that developed asking the traditional straightforward questions about sensitive behavior. In fact, in a study conducted by Droitcour et al. (1991), the estimates of engaging in sensitive behaviors developed using this methodology were actually lower than the estimates developed using the traditional straightforward questions. These data raised a number of questions about the value of the technique with extant methodology.

Exploratory Phase

For this study, several aspects of existing IC methodology were considered for modification, with the intent of potentially improving upon the technique. Characteristics of the survey were carefully taken into account in an effort to produce IC questions that were specifically tailored for the survey's content, instrumentation, and respondent pool.

Use of ACASI

It was decided that using audio computer-assisted self-interviewing (ACASI) offered several possibilities for decreasing error in the IC method. First, a sample task could be given to train the respondent on how to complete the task correctly. Second, if the respondent's answer was the item number of the most common item, ACASI could query the subject about this to determine if the respondent correctly understood the task. Finally, a variety of experiments could be introduced that would explain more about the type and degree of measurement error. Laboratory work would be needed to determine how to best train respondents on the task and what kinds of experiments would be the most useful.

Optimizing the Size of the List

At least two nonoverlapping lists of behaviors were suggested for implementing the IC method in the survey. For explanatory purposes, they are called "List 1" and "List 2." Both lists contain an equal number of nonstigmatizing behaviors until the "sensitive" behavior is added to one of the lists. Then the lists are referred to as the "short list" (sensitive behavior absent) and the "long list" (sensitive behavior present). Half the sample, call it "sample A," receives short List 1 and long List 2. The other half of the sample, "sample B," receives long List 1 and short List 2. Thus, four average counts are computed corresponding to the four lists: the average count for the two lists without sensitive behaviors and corresponding averages for the same two lists with the sensitive behavior added. These data yield two IC estimates of the prevalence of the sensitive behavior, one based upon the List 1 short/long pair and another based upon the List 2 short/long pair. A large random sample of the household population is needed to ensure that frequency of the less sensitive behaviors is similar across the two subsamples.

Cognitive laboratory studies of the IC methodology conducted by Droitcour et al. (1991) suggested that the number of behaviors in the list was a key determinant of response accuracy. If the lists are too short, respondents may fear that their privacy is not sufficiently protected (i.e., the other behaviors are not viewed as adequately "masking" the sensitive behavior). Consequently, deliberate underreporting of the sensitive behavior could result. Conversely, too many behaviors in the list may substantially increase the difficulty of the task because the participant has to consider more behaviors, determine their applicability, and tally the applicable number for a longer list. Droitcour et al. (1991) suggested that the optimal number of behaviors in the short list was between three and five.

Developing the Lists of Items

Prior research suggested that when the content of the list of behaviors was consistent with the topic of the questionnaire and/or the topic of the questionnaire section, respondents were less suspicious of the task, and this tended to increase truthful reporting. As an example, in the National Household Seroprevalence Survey pretest, lists for the IC methodology were developed that included behaviors that might put one at risk of contracting the acquired immunodeficiency syndrome (AIDS) virus (Droitcour et al., 1991). Because this was the topic of the study, respondents tended to be less suspicious of the IC methodology in that context. These findings suggested that more cognitive laboratory research was needed to determine which behaviors survey respondents would find the least threatening in the context of a drug use question. The behaviors would not need to be truly "innocuous" but should not be so sensitive that there would be a social desirability bias associated with reporting them. In fact, behaviors that are slightly counter to social norms might be seen as more consistent with drug use behaviors and thus less likely to arouse suspicion than completely innocuous ones. The goal of the research would be to develop a list of behaviors that would not have a substantial social desirability bias and that would be logically consistent with questions about drug use.

Explaining the Purpose of the Task

In previous research (e.g., Greenberg, Abul-Ela, Simmons, & Horvitz, 1969), because of the inclusion of a sensitive behavior among a list of innocuous ones, researchers have sometimes

offered complex explanations of why the respondent is being asked to do the task and how it will protect his or her privacy. However, as Wiseman, Moriaty, and Schafer (1975-1976) and Shimizu and Bonham (1978) showed, such explanations can backfire in that the respondent can become even more suspicious of the task. These explanations also increase respondent burden due to their complexity and the time they add to the interview. By developing contextually relevant lists of behaviors that include some behaviors that are mildly nonnormative, the survey could omit these complex explanations and make the task less burdensome and off-putting.

Wording of the Drug Item

Although not related strictly to the IC methodology, it was believed important to provide an unambiguous description of the drug-related behavior. Prior research (e.g., Biemer & Wiesen, 2002) on sensitive items suggested that when asked if they used a particular drug, respondents sometimes interpreted the question to mean they were a "user"—that is, they used the drug regularly or frequently. Asking instead an unambiguous question such as, "In the past 12 months, have you used cocaine powder at least once?" would be likely to increase the understanding of the question and so potentially increase reporting. These findings also suggested that the cognitive laboratory work should focus on the wording of the items as well as the topics covered by them.

Measurement Error and Bias

Droitcour et al. (1991) found sizable measurement errors and biases when using the IC methodology. However, it was considered possible to reduce these errors and improve the accuracy of the IC estimates by modifying the IC methodology as described above. Thus, an important objective of the research was to evaluate the measurement error in the IC methodology and to determine whether the careful consideration of the IC design features mentioned above would successfully reduce and control the effects of measurement error on the estimates.

Sample Size Issues

Under the assumption that measurement error could be controlled, a number of simulation studies were conducted to evaluate the standard errors (SEs) of the IC estimates under a range of assumptions about the prevalence of the item-level behaviors. These simulations suggested that behaviors having a very low (say less than 10 percent) or very high (say more than 90 percent prevalence) were ideal because the precision of the IC estimates was highest when the prevalence of the innocuous behaviors was near 0 or 1. Assuming that the prevalence of each behavior included on the short list was 10 percent or less, the minimum sample size required to achieve the desired level of precision was 35,000 responses. From a cognitive perspective, however, including behaviors that have a prevalence that is extremely high or low in the list was more likely to arouse suspicions of the task and/or be perceived to be nonmasking. For example, extremely low prevalence behaviors might make the question sound ludicrous and thus arouse suspicion. Extremely high prevalence behaviors might be seen to be nonmasking because the respondent assumes everyone will have engaged in the behavior. Therefore, behaviors having a frequency somewhat higher than 5 percent or lower than 95 percent might be required in order to make the task work effectively.

Conclusions from Exploratory Phase

Few options are available for increasing the accuracy of self-reported drug use in a survey. Promises of confidentiality and anonymity of reports can encourage accurate reporting, but these were already standard features of the survey's data collection methodology. Hair and urine tests are possible options, but these are costly and cumbersome and can be unreliable. The IC methodology is inexpensive to implement and has the potential to produce less-biased estimates. It also has the potential for increasing the acceptability of survey findings with NSDUH critics. But because of the problems described above with the method and the associated error, it was determined that the IC methodology, as it was used in the past, was not appropriate for estimating drug use prevalence. However, using a modified approach, a new IC module could be developed that would produce estimates potentially having greater statistical validity.

For these reasons, a program of research was mounted to develop a modified version of the IC methodology tailored to the needs of the survey. The aim of the modified approach was to design IC questions that minimize measurement errors and maximize estimator accuracy. The next section describes the research that was conducted to develop this modified approach and the resulting IC modules.

Implementation Phase

Start-Up

Many potential behaviors, as well as the format for the modules, were considered in the research. Contacts were made with researchers in the public health community soliciting their suggestions, and a number of instruments were reviewed from a variety of other studies. Websites with statistics on the frequency of various risk behaviors also were consulted, including ones for the following scales and instruments:

- Centers for Disease Control and Prevention (CDC) Youth Risk Behavior Surveillance System (YRBS);
- Temperament and Character Inventory or Tridimensional Personality Questionnaire (C. R. Cloninger);
- Sensation Seeking Scale (M. Zuckerman);
- Tension Risk Adventure Inventory (G. Keinan et al.);
- Arnett Inventory of Sensation Seeking;
- Barratt Impulsiveness Scale;
- the work of Hans Eysenck;
- the Hare Psychopathy Checklist Revised (PCL-R); and
- the National Highway Transportation Safety Administration (NHTSA) questionnaires about unsafe driving.

A literature search was conducted on risky behaviors to identify articles that might provide some innocuous behaviors for the IC questions. Some of the articles researched were "Physical Recklessness in Adolescence: Trait or Byproduct of Depressive/Suicidal States?" (Clark, Sommerfeldt, Schwarz, Hedeker, & Watel, 1990); "Adolescents' Perceptions of Their Risk-Taking Behavior" (Gonzalez et al., 1994); and "The Life Attitudes Schedule: A Scale to Assess Adolescent Life-Enhancing and Life-Threatening Behaviors" (Lewinsohn et al., 1995). Very few of these behaviors were found to have very low or very high frequency, were appropriately nonstigmatizing, and were able to be phrased in ways that were not overly complex. Therefore, the investigators had to invent risky behaviors that met these criteria. Approximately 50 behaviors that best met the constraints of the study were developed for testing in the first round. These constraints are discussed in detail in the following sections.

Placement of Module and Items

For youths, the IC module was placed in the behavior section following the drug use section because this section already asks about risk behaviors. This seemed to be the most appropriate position for the module because an IC question about various types of risky behaviors appearing in this section should seem consistent to the respondents and reduce any suspicion they may have about the motives of the question. For adults, the IC section was placed in the social environment section, which included questions about antisocial behaviors, including socially risky behaviors such as criminal behaviors.

Although not a necessary component of the IC approach, each behavior in the IC short list also was queried directly later in the questionnaire. The primary purpose of including a direct question for each behavior was to evaluate the measurement error in the IC questions. For example, consider a short-list IC question with four behaviors. If each of these behaviors also was queried directly, a second "pseudo-IC" response could be formed by counting the number of positive responses to the four individual questions for each respondent. Then, using test-retest methods, the reliability of the original IC question could be estimated. This information was essential for evaluating the measurement error in the IC methodology.

In addition, the inclusion of the direct questions provided data on the prevalence of the individual behaviors making up the short- and long-list questions as well as data on the inter-item correlation coefficients. These estimates will be quite useful for designing future IC studies. For example, in deriving the formulas for computing the power of the IC estimates, independence was assumed because no data on inter-item correlations for the behaviors in the lists were available. Although this may be a reasonable approximation, the sample size requirements could be better assessed with the inter-item correlation estimates provided by the direct questions.

Because the youth behaviors (for adolescents) and social environment (for adults) sections of the questionnaire ask specifically about rule-breaking and some risk-taking behaviors, and because these sections follow the risk behavior section, the questions asking about the individual items from the short lists were placed in these sections.

Format for Round 1 Cognitive Testing

Two rounds of cognitive testing were conducted. The first round used a paper-and-pencil interviewing (PAPI) instrument rather than ACASI, even though the final instrument would be implemented in ACASI. This was done in order to meet the schedule for implementing the IC modules in the survey. Following this first round of testing, the modules were revised and converted to the ACASI mode, which then was tested in the second round.

The first round of testing incorporated the following design features:

- A simple format with no reasons given for the task because Droitcour et al. (1991) found that explaining the purpose of the task was both confusing and also aroused suspicion among respondents.
- A "sample" or "practice" list before the real lists because (1) the first list could serve as a training tool so that if the respondent made errors on the first list, he or she could repeat the task until he or she was "trained" in it and (2) higher frequency behaviors could be included in the first "practice" list. This increased the likelihood that the respondent would say "yes" to at least one item and become comfortable with responding "yes" to list items.
- A three-item short list and a four-item long list (i.e., the list including cocaine use) because this was the smallest number of items thought to provide some anonymity, better statistical properties, and less respondent burden than longer lists.

Choice of Items

A number of criteria were developed for determining the best behaviors to include as items for the short list. In the first round of testing, these criteria were the following:

- As previously discussed in the section on sample size issues, behaviors had to be very low frequency (under 10 percent) or very high frequency (over 90 percent). Because information was difficult to obtain on the exact probabilities of many of these behaviors, those with a prevalence rate of 5 percent or less or 95 percent or more were preferred. This approach would allow a margin of error.
- Short-list items had to be "risky" behaviors. It was considered important for the respondents' acceptance of the task that these lists be perceived as "risky behavior scales" even though they were not described as such. Therefore, items such as "was diagnosed with a serious illness" or "was hospitalized overnight," which are not risky behaviors, were not included. The testing supported this approach, as participants were uncomfortable when items on the list were events that may not have been the respondent's "fault," that is, not volitional. For example, participants were uncomfortable with an item asking if they had been in a car accident in which someone was killed because such an accident may not have been the result of engaging in any risk behavior—a drunk driver might have hit them when they were driving safely.

- Respondents had to recognize the behaviors as risky. Therefore, many behaviors, such as not using sunscreen/sunblock or eating high-fat foods, were excluded because certain groups of people were unlikely to perceive the behaviors to be a serious risk to them.
- Behavior descriptions could not be too complex or too wordy. This was sometimes difficult because often what made the behavior low frequency was the conditions under which it took place or the extreme nature of the event.
- Behaviors had to be perceived by respondents to not be outrageous or incredible but merely risky and uncommon.
- Drug or alcohol risk behaviors that had been asked about previously in the questionnaire could not be used because of possible contextual effects.
- Short-list items could not be highly stigmatizing or the purpose of the task would be defeated. If cocaine was just one of several very stigmatizing behaviors, respondents would be unlikely to admit engaging in any of them. Thus, many potential items related to illegal or other very stigmatizing behaviors were eliminated.
- In order to effectively mask admitting to cocaine use, behaviors could not be extremely or almost totally correlated with age or any other sociodemographic characteristic. For example, if a behavior was likely to be perceived to have essentially a zero probability among adults over 65 (e.g., bungee jumping), it was not included because respondents older than 65 responding "yes" to just one item were likely to think that no one would believe that the item they did was bungee jumping.
- Although many other behaviors (e.g., trying to stop a fight or an assault) were more likely to be exhibited by younger males than individuals in other sociodemographic groups, these items also were not implausible for women and older men in that the person might be put into circumstances where he or she had little choice but to engage in the risky behavior. For example, a mother or grandmother might try to stop her children or grandchildren from fighting. She also might be forced to walk through a dangerous neighborhood at night. In general, elderly women are the least likely to engage in many risk behaviors, so it was difficult to find items that were not totally implausible for them. In general, recreational behaviors that were associated with a younger age group were avoided (e.g., motorcycle riding, rock climbing, and bungee jumping). In many cases, however, behaviors that were merely unlikely for this group of individuals but not impossible were still retained because circumstances might bring about their engaging in the behavior.

Round 1 Cognitive Testing

Figure 9.1 lists the behaviors that were tested in the first round and the final resolution on each item based on the results from the *first* round of testing. Both sample/practice items that are used to help train the respondent were included, as well as items for the short lists that are used to mask the reporting of the sensitive behavior.

Figure 9.1 Items Tested in Round 1

Sample/Practice Items Tested in Round 1

- Hitchhiked [DROPPED—TOO CORRELATED WITH AGE]
- Crossed railroad tracks when a train was coming and almost got hit by the train [KEPT BUT LITERATURE SEARCH SUGGESTED INFREQUENT ENOUGH TO INCLUDE AS MASKING ITEM]
- Went someplace where you knew you might get shot, stabbed, or beaten up [KEPT BUT REWORDED TO ADDRESS RESPONDENTS' CONCERNS].

Masking Items Tested in Round 1

- Went swimming during a thunderstorm [REWORDED TO MAKE MORE PLAUSIBLE/FREQUENT]
- Drove a motorcycle on the highway without wearing a helmet [DROPPED—TOO CORRELATED WITH AGE/GENDER]
- Went bungee jumping [DROPPED—TOO CORRELATED WITH AGE]
- Jumped out of a moving car [DROPPED—RESPONDENTS SAW AS TOO IMPLAUSIBLE]
- Shot a gun inside a house [DROPPED—RESPONDENTS SAW AS TOO IMPLAUSIBLE]
- (Adult) drove a car more than 90 mph [CHANGED TO 100 MPH BASED ON DATA AND RESPONDENTS' REPORTS OF A LOT OF PEOPLE DOING THIS ON INTERSTATES WITH SPEED LIMITS OF 75 MPH]
- (Adolescent) ran away from home and slept on the street [KEPT AS WORDED]
- Took a gun away from an angry person [DROPPED—RESPONDENTS SAW AS TOO IMPLAUSIBLE]
- Jumped out of a second-story window [DROPPED—RESPONDENTS SAW AS TOO IMPLAUSIBLE]
- Was in a car accident in which someone was killed [DROPPED—RESPONDENTS UNCOMFORTABLE WITH ITEM BECAUSE SEEN AS NONVOLITIONAL]
- Was a witness to a violent crime [DROPPED—RESPONDENTS UNCOMFORTABLE WITH ITEM BECAUSE SEEN AS NONVOLITIONAL]
- Borrowed money from a loan shark [REWORDED—RESPONDENTS SAW "LOAN SHARK" AS POOR TERM]
- Rode with a drunk driver [KEPT AS WORDED BUT AS SAMPLE ITEM BECAUSE OF PROBABLE FREQUENCY].

Sensitive Item

- Used cocaine, in any form, one or more times [KEPT AS WORDED].

Nine subjects were tested in Round 1: (a) five white females—one in her 20s, three in their 30s, and one aged 70; (b) two black females in their 30s; (c) one white male in his 40s; and (d) one black male in his 20s. The black male and one of the black females were cocaine users in substance abuse treatment.

Of the nine respondents, only one showed any real problem completing the task. A male cocaine user tried to circle the item(s) instead of going on to where he was asked the number of items. (Recall that Round 1 used PAPI.) Using ACASI will help to eliminate this problem because it will not allow subjects to circle the items. The letter preceding each item on the lists in Round 1 was also removed to de-emphasize them so that they are not selected individually by the respondent. Two of the women had a little difficulty when they tried the first list but were able to do the task correctly. After the first list, they understood the task. This tends to support the continued use of a sample list.

Participants appeared to perceive the lists to be lists of "dangerous" things one might do. There did not appear to be any suspicion about the task in general. However, some items raised questions with the respondents. For example, "crossing the tracks when a train was coming and almost getting hit" seemed more reasonable to most participants than items like "jumping out of moving vehicles," which were viewed as "outrageous" or "crazy" and, therefore, dropped.

A number of items seemed plausible to some participants and implausible to others. Most strikingly, the cocaine abusers found some items plausible—like shooting a gun in the house—that nonabusers found quite outrageous. Those items were eliminated from the list because most respondents were likely to find them implausible. The objective here was to find behaviors, such as riding with a drunk driver, that people from all strata of society would find plausible, though uncommon and risky.

In reviewing the list of risky behavior items, some participants were uncomfortable if the behavior/event was not wholly volitional. For example, participants found "being in a car accident in which someone was killed" or "being a witness to a violent crime" to be cognitively dissonant in this framework. Although these are certainly risky experiences, because the accident or the witnessing might not have been under the control of the respondent, they did not feel it appropriate to include such items in the lists. Items that were likely to be perceived as not volitional were dropped.

Participants felt that a longer list would make them feel more comfortable admitting to a sensitive behavior such as cocaine use, but they varied in their appraisal of the importance of lengthening the lists. Some explicitly stated that a list of four items was too short to camouflage their answers, especially because the other items on the list were rare events.

On the basis of this research, the second round of testing focused on IC modules that incorporated a four-item (short) and five-item (long) list. A six-item list also was tested, although the research team was concerned that a six-item (long) list is too long to be answered accurately. Although six items may not be a serious obstacle for good readers, for poor readers (or nonreaders) it could be an issue. For example, if the module is too long, the module presents a memory task because non- or poor readers cannot use the items on the screen as a reference. For these reasons, six items may be too many for people to remember and evaluate, even with the

option of having the list repeated in ACASI. Few respondents are likely to spend too much time repeating the list whether or not they can remember all of the items well.

Based on the results from the first nine participants, the module was modified before it was tested on ACASI. As noted above, items with serious problems were replaced, and the list length was expanded by one item.

Round 2 Cognitive Testing

Figure 9.2 shows the items that were tested in the second round and the final resolution on each based on the results of the *second* round of testing.

Seven subjects were tested in Round 2: (a) three white females—one each in their 30s, 40s, and 60s; (b) one white male in his 60s; (c) two adolescent white males around ages 15 to 17; and (d) one black male in his 30s. The black male was a client in substance abuse treatment.

No participant showed any suspicion of the task or had any real problem completing the task. One adult male participant with a lower education level seemed to feel he had to listen to everything carefully (which he did) but did the task satisfactorily. Most participants found the items to be related and recognized them as risky or dangerous. Probably because there were several driving questions, a couple of participants thought the purpose of the questions was to determine whether they were unsafe drivers. Participants generally felt that they would feel reasonably comfortable admitting to a sensitive behavior in a list rather than reporting it individually. The four (without cocaine) and five (with cocaine) item lists seemed to be long enough to make participants reasonably comfortable in answering truthfully, although, as with the Round 1 participants, most felt that the longer the list, the more their answers would be masked. A couple of the masking items were troublesome: Russian roulette was seen as both jarring and something that some adolescents might not understand, and traveling to another country where there was a war seemed too implausible. These items were both dropped.

Adolescent participants were no different from adults; that is, they also were not suspicious of the task, had no difficulty completing the task, and had no difficulty understanding the individual items. None of the participants found the verification of their answers to be very troublesome. In fact, one adolescent said it made him feel more comfortable to know he had not actually hit a wrong key.

Wording of Items in the Final Module

The process for choosing the types of behaviors to include in the module was described above. The individuals involved in developing the module have all had previous experience in developing and testing items for questionnaires and have a good understanding of the characteristics of a well-worded question. Nonetheless, prior to conducting cognitive testing of the questions, the wording and structure of each question were reviewed by experts in question construction. Cognitive testing of the revised questions then was conducted to ascertain not only what questions did not work well but why they did not work well. In general, most questions were well worded, and respondents both understood the items and were able to answer them. If the behavior/event was viewed as too implausible, it was dropped. But if the item was seen as

Figure 9.2 Items Tested in Round 2

Sample/Practice Items in Round 2

- Rode with a drunk driver [KEPT AS WORDED]
- Walked alone after dark through a dangerous neighborhood [KEPT AS WORDED]
- Rode a bicycle without a helmet [KEPT AS WORDED]
- Went swimming or played outdoor sports when it was lightning [KEPT AS WORDED].

Masking Items Tested in Round 2

- (Adult) drove a car more than 100 miles per hour [KEPT AS WORDED]
- (Adolescent) ran away from home and slept on the street [KEPT AS WORDED]
- Borrowed money from someone who might injure you if you didn't pay it back [KEPT AS WORDED]
- Smoked more than three packs of cigarettes in a day [DROPPED—CONCERN ABOUT NOT MASKING FOR THOSE PREVIOUSLY REPORTING NOT SMOKING IN THE PAST YEAR]
- Used steroids to become more muscular [KEPT AS WORDED]
- Crossed railroad tracks when a train was coming and almost got hit by the train [KEPT AS WORDED]
- Was seriously injured from fighting [KEPT AS ADOLESCENT ITEM]
- Took a knife or gun away from an angry person [REWORDED AS AN ADULT ITEM. PLAUSIBILITY OF ITEM WAS INCREASED BY CHANGING ITEM FROM STOPPING SOMEONE USING A KNIFE OR GUN TO STOPPING SOMEONE FROM FIGHTING OR ASSAULTING SOMEONE. FREQUENCY WAS KEPT DOWN TO ACCEPTABLE RANGE BY ADDING THAT PARTICIPANT WAS INJURED IN THIS EVENT.]
- Used laxatives or vomited on purpose in order to keep your weight down [KEPT AS WORDED]
- Used cocaine, in any form, one or more times [KEPT AS WORDED]
- (Adult) passed another vehicle when you knew it was not safe to pass [KEPT AS WORDED]
- (Adolescent) hacked into a government computer system [KEPT AS WORDED]
- Went to a foreign country where a war was going on [DROPPED BECAUSE RESPONDENTS SAW AS TOO IMPLAUSIBLE]
- Played Russian roulette [DROPPED BECAUSE SOME RESPONDENTS WERE UNFAMILIAR WITH THE TERM]
- Was careless and set a large or serious fire with a cigarette [KEPT AS WORDED]
- Gained or lost more than 50 pounds [KEPT AS WORDED].

implausible only because it was too extreme, in some cases the restrictions were relaxed to make the item more plausible. For example, an item that asked about stopping a knife or gun attack was reworded to "stopping a fight."

If the participants just had problems with the wording of an item, however, the item was reworded. For example, in Round 1, participants were not comfortable with the term "loan shark." It was viewed as suggesting gangsters and other things normal people do not have in their lives. The items were reworded so that "loan shark" was replaced by "someone who might injure you if you didn't pay back the money you borrowed from them." Most participants in Round 2 did not seem to find this item troublesome, although one did state that she thought that this was a very low frequency event.

At the end of Round 2 of cognitive testing, the items seen as most problematic were dropped. However, there were still items that, given more time, could have benefited from further testing and/or development. These included the following:

- Borrowed money from someone who might injure you if you didn't pay it back.

Only one participant found this highly implausible, but it would have been good to test this out on other individuals.

- Took a knife or gun away from an angry person/were injured when you tried to stop a fight or an assault.

In Round 2, the item "was seriously injured from fighting" was meant to be asked only of adolescents, but it was mistakenly asked for adults as well. There was a concern that this item would not be sufficiently masking for older women and men. "Took a knife or gun away from an angry person" had been asked of adults, but it was seen as too implausible to include as the adult substitution for the adolescent fighting item, so the item was reworded to "were injured when you tried to stop a fight or an assault." This seemed to be more plausible across age/gender groups. Even an older woman, for example, might try to stop her children or grandchildren from fighting or might try to stop her husband from assaulting a child. However, this item has had no cognitive testing.

- (Adult) passed another vehicle when you knew it was not safe to pass.

Some individuals were not certain what was included as "not safe to pass." Although they answered the question, it could still be improved.

- (Adolescent) hacked into a government computer system.

This item was tested on only two adolescents. It would have been useful to do more testing of this item.

Based on the overall findings, it seemed feasible to administer the IC methodology as part of the survey's ACASI interview with little or no problems for most respondents. The final modules that were developed are shown in ***Figures 9.3*** and ***9.4***. ***Figure 9.3*** shows the practice question

that was administered to all respondents regardless of age or random sample assignment. As described previously, a practice sequence was included (1) to serve as a training tool where the respondent could make errors and repeat the sequence until he or she was "trained" in the task, and (2) to include an illustration with higher frequency items so that the respondent would have an opportunity to respond "yes" to at least one item and thereby become comfortable with responding "yes" to list items. ***Figure 9.3*** also shows the verification questions that were used following each item count question.

Figure 9.4 shows the youth and adult IC questions that were administered for Random Sample A. The questions for Random Sample B are not shown because they were identical to those for Random Sample A except that the cocaine item was added to the ICQ1(S) questions and deleted from the ICQ2(L) questions. The wording of the IC question and the verification questions was identical to that shown in ***Figure 9.3*** with the items in ***Figure 9.4*** substituted. These items were selected because, in the pretest work, they seemed to work well with whites, blacks, adolescents, drug users, and respondents of different ages and from different backgrounds. They were not tested with Hispanics nor with those with reading difficulties.

Sample Size Calculation

The sample size requirements for the IC implementation in the survey, assuming a four-item short list (or, equivalently, a five-item long list), were estimated using simulation. The purpose of this analysis was to ensure that there would be adequate precision in the IC estimates to test hypotheses about the prevalence of cocaine use in the United States for various population subgroups. In this section, some details regarding these calculations are provided.

The single-question IC estimator is

$$\bar{x}_{diff} = \bar{x}_L - \bar{x}_S , \tag{1}$$

where $\bar{x}_L$ is the mean number of items selected from the long list of items and $\bar{x}_S$ is the mean number of items from the short list. Assume that the sample consisting of $2n$ households is split into halves, with half receiving the long list and the other half the short list. The variance of this estimator is

$$\mathrm{Var}(\bar{x}_{diff}) = \sigma n^{-1}(\sigma_L^2 + \sigma_S^2), \tag{2}$$

where σ_L^2 is the variance of a single long-list response, σ_S^2 is the variance of a single short-list response, and δ is the design effect.

Of particular interest in this study was the estimator for a double-list design. In a double-list design, the sample is split into subsamples of approximately equal size, which are referred to as Sample A and Sample B. Sample A received the short-list IC question, denoted by ICQ1(S), and Sample B received the long-list version of this question, denoted ICQ1(L). Recall that ICQ1(L) includes all items in ICQ1(S) with one additional item: the past year cocaine use item. To complete the symmetry of the design, two additional questions are added, ICQ2(S) and ICQ2(L), where ICQ2(S) is a different short-list question administered to Sample B respondents while

Figure 9.3 Item Count Practice Question with Verification

This practice series was administered to all respondents regardless of age or sample assignment. A similar series of questions was used for all item count questions.

1. Here is a list of things that you may or may not have done during the past 12 months:

Rode with a drunk driver

Walked alone after dark through a dangerous neighborhood

Rode a bicycle without a helmet

Went swimming or played outdoor sports during a lightning storm

How many of the things on this list did you do during the past 12 months, that is, since [DATE FILL]?

0 None of these things

1 One of these things

2 Two of these things

3 Three of these things

4 All four of these things

2. The computer recorded that you did [FILL FROM 1] from the list below during the past 12 months:

Rode with a drunk driver

Walked alone after dark through a dangerous neighborhood

Rode a bicycle without a helmet

Went swimming or played outdoor sports during a lightning storm

Is that correct?

1 Yes

2 No

3. [IF RESPONSE TO 2 IS NO] Please answer this question again. How many of the things on this list did you do during the past 12 months, that is, since [DATE FILL]?

Rode with a drunk driver

Walked alone after dark through a dangerous neighborhood

Rode a bicycle without a helmet

Went swimming or played outdoor sports during a lightning storm

0 None of these things

1 One of these things

2 Two of these things

3 Three of these things

4 All four of these things

Figure 9.4 Youth and Adult Items Used for Random Sample A

These items were used in the item count questions like those shown in Figure 9.3 for Random Sample A. For Random Sample B, the questions were identical except the cocaine item was deleted from ICQ2(L) and added to ICQ1(S) to form ICQ2(S) and ICQ1(l), respectively.

Youth Modules	**Adult Modules**
ICQ1(S) Items	ICQ1(S) Items
Ran away from home and slept on the street	Drove a car more than 100 miles per hour
Gained or lost more than 50 pounds	Gained or lost more than 50 pounds
Were seriously injured in a fight	Were injured when you tried to stop a fight or an assault
Used steroids to become more muscular	Used steroids to become more muscular
ICQ2(L) Items	ICQ2(L) Items
Crossed railroad tracks when a train was coming and almost got hit by the train	Crossed railroad tracks when a train was coming and almost got hit by the train
Used laxatives or vomited on purpose in order to keep your weight down	Used laxatives or vomited on purpose in order to keep your weight down
Hacked into a government computer system	Passed another vehicle when you knew it was not safe to pass
Used cocaine, in any form, one or more times	Used cocaine, in any form, one or more times
Was careless and set a large or serious fire with a cigarette or a match	Was careless and set a large or serious fire with a cigarette or a match

ICQ2(L) is the long-list counterpart of that question administered to Sample A respondents. ***Figure 9.5*** summarizes this design.

Figure 9.5 Design of the Item Count Experiment

	Sample A	**Sample B**
Short-List Question	ICQ1(S)	ICQ2(S)
Long-List Question	ICQ2(L)	ICQ1(L)

Then the estimator of cocaine use from List 1 can be written as

$$\hat{P}_1 = \bar{x}_{L(1)} - \hat{x}_{S(1)} \tag{3}$$

and for List 2 as

$$. \hat{p}_2 = \bar{x}_{L(2)} - \hat{x}_{S(2)} . \tag{4}$$

Combining these two estimators, the IC estimate for the entire sample is

$$\hat{p} = (\hat{p}_1 + \hat{p}_2)/2. \tag{5}$$

The variance of the combined estimator is

$$\mathrm{Var}(\hat{p}) = 0.25[\mathrm{Var}(\hat{p}_1) + \mathrm{Var}(\hat{p}_2) - 2\rho_{12}\sqrt{\mathrm{Var}(p_1)\mathrm{Var}(p_2)}] \tag{6}$$

where ρ_{12} is the correlation between the estimators $\hat{p}_1$ and $\hat{p}_2$. It can be shown that $\rho_{12} = -c_1\rho_1 - c_2\rho_2$, where c_1 and c_2 are non-negative constants, ρ_1 is the correlation between $\bar{x}_{L1}$ and $\bar{x}_{S2}$ and ρ_2 is the correlation between $\bar{x}_{L2}$ and $\bar{x}_{S1}$. Assuming ρ_1 and ρ_2 are non-negative, which is expected in almost all practical applications, ρ_{12} will be negative. Thus, a conservative expression for $\mathrm{Var}(\hat{p})$ is $0.25[\mathrm{Var}(\hat{p}_1) + \mathrm{Var}(\hat{p}_2)]$.

The variance in equation (6) is a function of the response probabilities associated with the items in two pairs of lists. To see this, note that the variance of a single response to the long list is given by

$$\sigma_L^2 = \sum_{k=0}^{5} k^2 Pr(X_L = k) - [\sum_{k=0}^{5} kPr(X_L = k)]^2, \tag{7}$$

where $Pr(X_L = k)$ is the probability $X_L = k$ for $k = 0,...,5$. Let I_j be an indicator variable corresponding to the true status of a randomly selected respondent to item j on the long list for $j = 1, ..., 5$ (i.e., $I_j = 1$ if the true status is positive and $I_j = 0$ if the true status is negative). Then $\Pr(I_j = 1) = p_j$ is the prevalence of item j. For each individual, j, there is a combination of I_1, I_2, I_3, I_4, and I_5 reflecting the individual's true status on all five items. The probability of a particular combination, (I_1, I_2, I_3, I_4, I_5), is

$$\Pr(I_1, I_2, I_3, I_4, I_5) = \prod_{\mathrm{j}} p_j^{I_j}(1-p_j)^{1-I_j} + \gamma(I_1, I_2, I_3, I_4, I_5), \tag{8}$$

where γ (I_1, I_2, I_3, I_4, I_5) includes terms for the joint probabilities associated with the combination. Now $Pr(X_L = k)$ can be written as

$$Pr(X_L = k) = \sum_{(I_1,...,I_5) \in S_k} Pr(I_1, I_2, I_3, I_4, I_5) \tag{9}$$

and $S_k = \{(I_1,..,I_5) | \sum_{j=1}^{5} I_j = k\}$; i.e., the sum is over all combinations of (I_1, I_2, I_3, I_4, I_5) that result in $\sum_{j=0}^{5} I_j = 5$. A similar form can be derived for the variance of a short-form response denoted σ_S^2.

The magnitude of γ (I_1, I_2, I_3, I_4, I_5) in equation (8) is unknown; however, assuming γ (I_1, I_2, I_3, I_4, I_5) = 0 (independence between the items in a list) is likely to result in understating σ_L^2 and σ_S^2 because in most practical situations correlations between the items are positive. Thus, the sample size requirements based upon this assumption also will be likely somewhat understated. It also is assumed that ρ_{12} in equation (6) is 0, which, as previously noted, is likely to overestimate the variance. The effects of these two assumptions may be somewhat offsetting. Therefore, equation (6) with $\rho_{12} = 0$ and γ (I_1, I_2, I_3, I_4, I_5) = 0 should provide a reasonable approximation of the variance of the double-list estimator for the purposes of estimating the power associated with the various design options. However, the variance of the single-list estimator using equation (4) is likely to be slightly underestimated by this approach.

To determine $\text{Var}(\hat{p})$ for a given sample size, *n*, consider three cases for the prevalence of the innocuous items in the lists as follows:

- Case 1: $p_j = 0.05$ for $j = 1, ..., 4$ (i.e., the prevalence of the innocuous items are all set to 5 percent).
- Case 2: $p_j = 0.10$ for $j = 1, ..., 4$ (i.e., the prevalence of the innocuous items are all set to 10 percent).
- Case 3: $p_j = 0.20$ for $j = 1, ..., 4$ (i.e., the prevalence of the innocuous items are all set to 20 percent).

The expected value of the IC estimator for cocaine was assumed to be 0.03. The coefficient of variation of the estimate is the square root of $0.5\,\text{Var}(\hat{p})/0.03$ for a double-list design.

The sample size calculations were derived from the consideration of the power of the hypothesis test: H_0: $\pi = \pi_0$ versus the alternative H_a: $\pi > \pi_0$, where π is the true prevalence of past year cocaine use and π_0 is a specified constant. Statistical power is the probability this one-sided test is rejected when the true cocaine prevalence rate is 3 percent for specified values of π_0. ***Table 9.1*** shows the calculations for *n* ranging from 1,000 to 35,000 households for both the single-list design and the double-list design for $\pi_0 = 0.015$.

The maximum sample size available for the IC experiment is $2n = 70{,}000$ households, and thus the maximum *n* in the sample size calculations is 35,000. Assuming that a power of at least 80 percent is desired, then either the single- or the double-list design will achieve the minimally acceptable power with the available sample size if the prevalence of the innocuous items is as small as 5 percent. At 10 or 20 percent, only the double-list design will provide acceptable power.

Table 9.1 Power Analysis for Item Count Estimation of Cocaine Prevalence

	Innocuous Item Prevalence = 5				Innocuous Item Prevalence = 10				Innocuous Item Prevalence = 20			
	Double List		Single List		Double List		Single List		Double List		Single List	
n	C.V.	Power	C.V.	Power	C.V.	Power	C.V.	Power	C.V.	Power	C.V.	Power
1,000	65.9	18.6	93.2	13.3	87.4	14.0	123.7	10.6	114.4	11.3	161.8	9.0
2,000	46.6	28.2	65.9	18.6	61.8	20.0	87.4	14.0	80.9	15.1	114.4	11.3
3,000	38.0	36.9	53.8	23.6	50.5	25.5	71.4	17.1	66.1	18.6	93.4	13.2
4,000	32.9	44.7	46.6	28.2	43.7	30.6	61.8	20.0	57.2	21.9	80.9	15.1
5,000	29.5	51.9	41.7	32.6	39.1	35.5	55.3	22.8	51.2	25.1	72.4	16.9
6,000	26.9	58.3	38.0	36.9	35.7	40.2	50.5	25.5	46.7	28.1	66.1	18.6
7,000	24.9	64.0	35.2	40.9	33.1	44.5	46.7	28.1	43.2	31.1	61.2	20.3
8,000	23.3	69.0	32.9	44.7	30.9	48.7	43.7	30.6	40.5	33.9	57.2	21.9
9,000	22.0	73.5	31.1	48.4	29.1	52.6	41.2	33.1	38.1	36.7	53.9	23.5
10,000	20.8	77.3	29.5	51.9	27.7	56.3	39.1	35.5	36.2	39.4	51.2	25.1
11,000	19.9	80.7	28.1	55.2	26.4	59.7	37.3	37.9	34.5	42.0	48.8	26.6
12,000	19.0	83.6	26.9	58.3	25.2	63.0	35.7	40.2	33.0	44.6	46.7	28.1
13,000	18.3	86.1	25.8	61.2	24.3	66.0	34.3	42.4	31.7	47.0	44.9	29.6
14,000	17.6	88.3	24.9	64.0	23.4	68.8	33.1	44.5	30.6	49.4	43.2	31.1
15,000	17.0	90.1	24.1	66.6	22.6	71.4	31.9	46.6	29.5	51.7	41.8	32.5
16,000	16.5	91.7	23.3	69.0	21.9	73.8	30.9	48.7	28.6	53.9	40.5	33.9
17,000	16.0	93.0	22.6	71.3	21.2	76.0	30.0	50.7	27.8	56.0	39.2	35.3
18,000	15.5	94.2	22.0	73.5	20.6	78.1	29.1	52.6	27.0	58.1	38.1	36.7
19,000	15.1	95.1	21.4	75.5	20.1	80.0	28.4	54.5	26.2	60.1	37.1	38.1
20,000	14.7	95.9	20.8	77.3	19.6	81.8	27.7	56.3	25.6	62.0	36.2	39.4
21,000	14.4	96.6	20.3	79.1	19.1	83.4	27.0	58.0	25.0	63.8	35.3	40.8
22,000	14.0	97.2	19.9	80.7	18.6	84.9	26.4	59.7	24.4	65.5	34.5	42.0
23,000	13.7	97.7	19.4	82.2	18.2	86.3	25.8	61.4	23.9	67.2	33.7	43.3
24,000	13.4	98.1	19.0	83.6	17.8	87.5	25.2	63.0	23.4	68.8	33.0	44.6
25,000	13.2	98.4	18.6	84.9	17.5	88.7	24.7	64.5	22.9	70.4	32.4	45.8
26,000	12.9	98.7	18.3	86.1	17.1	89.7	24.3	66.0	22.4	71.8	31.7	47.0
27,000	12.7	98.9	17.9	87.2	16.8	90.7	23.8	67.4	22.0	73.3	31.1	48.2
28,000	12.5	99.1	17.6	88.3	16.5	91.6	23.4	68.8	21.6	74.6	30.6	49.4
29,000	12.2	99.3	17.3	89.2	16.2	92.4	23.0	70.1	21.2	75.9	30.0	50.6
30,000	12.0	99.4	17.0	90.1	16.0	93.1	22.6	71.4	20.9	77.1	29.5	51.7
31,000	11.8	99.5	16.7	91.0	15.7	93.7	22.2	72.6	20.5	78.3	29.1	52.8
32,000	11.6	99.6	16.5	91.7	15.5	94.3	21.9	73.8	20.2	79.4	28.6	53.9
33,000	11.5	99.7	16.2	92.4	15.2	94.9	21.5	74.9	19.9	80.5	28.2	55.0
34,000	11.3	99.7	16.0	93.0	15.0	95.4	21.2	76.0	19.6	81.5	27.8	56.0
35,000	**11.1**	**99.8**	**15.8**	**93.6**	**14.8**	**95.8**	**20.9**	**77.1**	**19.3**	**82.5**	**27.4**	**57.1**
36,000	11.0	99.8	15.5	94.2	14.6	96.3	20.6	78.1	19.1	83.4	27.0	58.1
37,000	10.8	99.8	15.3	94.7	14.4	96.6	20.3	79.1	18.8	84.3	26.6	59.1
38,000	10.7	99.9	15.1	95.1	14.2	97.0	20.1	80.0	18.6	85.2	26.2	60.1
39,000	10.6	99.9	14.9	95.6	14.0	97.3	19.8	80.9	18.3	86.0	25.9	61.0
40,000	10.4	99.9	14.7	95.9	13.8	97.5	19.6	81.8	18.1	86.7	25.6	62.0

Notes:
n = number of cases receiving sensitive item
C.V. = coefficient of variation
Cocaine prevalence (π_0) = 0.03
Item independence assumed
Power computed at $\pi = 0.015$
Split half sample between sensitive and innocuous items
Number of innocuous items = 4

Further, if subnational estimates and estimates of cocaine use by various analytic domains are of interest, other rows of the table are relevant. For example, assuming an average prevalence of innocuous items of 10 percent and a double-list design, the smallest domain for which the power of the test of H_0 will still be acceptable is about 50 percent (i.e., 0.5 n). Based upon these calculations, it was recommended that the double-list design be used with an average prevalence of innocuous items of 10 percent or less and a sample size of 35,000 respondents receiving each list.

Data Collection Results

Of the 171,519 eligible households sampled for the 2001 NSDUH main study, 157,471 were successfully screened for a weighted screening response rate of 91.9 percent. In these screened households, a total of 89,745 sample persons were selected, and completed interviews were obtained from 68,929 of these sample persons, for a weighted interview response rate of 73.3 percent. The overall weighted response rate, defined as the product of the weighted screening response rate and weighted interview response rate, was 67.3 percent. The cumulative item nonresponse for the variables used in the analysis reduced the sample size in the analysis to 68,305 persons.

Analysis Phase

This section summarizes the results of a preliminary analysis of data collected for the IC experiment. IC estimates of the proportion of the population who have used cocaine in the past 12 months are presented, as well as data that will aid in the evaluation of the accuracy of these estimates. Additional analyses will be described that are currently being conducted with these data and that could yield improved estimates of past year cocaine use.

Item Count Estimates of Past Year Cocaine Use

Both weighted and unweighted versions of the IC estimator in equation (5) were computed in order to assess the effects of weighting on the estimates. Weighted estimates often increase the SEs of estimates, so it would not be surprising that the weighted IC estimates are more unstable than the unweighted ones.

One additional feature of the implementation of the previously described IC design was a verification question that provided an opportunity for respondents to correct their responses to the IC questions (see ***Figure 9.3***). Both their initial responses and their verified or corrected responses were recorded. This permits the calculation of IC estimates from the data as originally entered as well as after the verification question. If the verification approach was successful at reducing measurement error, the IC estimates based upon corrected data should be more accurate.

In ***Table 9.2***, the weighted and unweighted IC estimates of past year cocaine use based upon the corrected data (i.e., after verification by the respondent) are presented for age by gender groups. For comparison purposes, the 2001 survey estimates also are shown. Recall that the 12 to 17 age group received IC short lists that were slightly different from those for the 18 or older group and that were more appropriate for an adolescent population. ***Table 9.3*** shows the corresponding estimates of past year cocaine use based upon the uncorrected (or unverified) data.

Table 9.2 Unweighted and Weighted Item Count Estimates of Past Year Cocaine Use Prevalence, by Age and Gender—After Verification

Age	Gender	Unweighted IC	Weighted IC	2001 Survey
12 to 17	Total	0.71	0.73	1.5
	Male	0.14	0.19	1.4
	Female	1.29	1.28	1.5
18 or Older	Total	0.59	-0.08	1.9
	Male	1.43	0.42	2.8
	Female	-0.19	-0.55	1.1

IC = item count.

A surprising finding from ***Table 9.2*** is that all of the IC estimates of past year cocaine use are smaller than the survey estimates based upon direct questioning. This is unexpected because by design the IC methodology is intended to produce estimates that are greater than or equal to the estimates from direct self-reports. In fact, as discussed earlier, they should be substantially larger because estimates based upon direct self-reporting are likely to be lower than actual prevalence. Apparently, the efforts to develop IC modules that minimized the nonsampling error usually associated with this methodology were not successful.

The comparison of the verified and unverified estimates in ***Tables 9.2*** and ***9.3***, respectively, indicate the verification approach increased the estimates somewhat. Had the data not been verified, the IC estimates would be even more implausible. Four of the weighted IC estimates are negative in ***Table 9.3***, whereas only two estimates are negative in ***Table 9.2***. IC estimates will be biased downward if true cocaine users tend not to count cocaine use in their IC reports. Because cocaine use is a highly stigmatized, illegal behavior, underreporting of cocaine use can be a problem for IC questioning just as it is for direct questioning. The fact that the verified data estimates are larger than original data estimates provides some evidence that IC estimates of cocaine use, at least in this application, are negatively biased. The results in the table suggest that the verification questions reduced the negative bias in the IC estimates to some extent, but it is evident from the comparison of these estimates with the survey estimates that a substantial amount of bias still remains.

Table 9.3 Unweighted and Weighted Item Count Estimates of Past Year Cocaine Use Prevalence, by Age and Gender—Before Verification

Age	Gender	Unweighted IC	Weighted IC	2001 Survey
12 to 17	Total	0.07	0.05	1.5
	Male	-0.66	-0.61	1.4
	Female	0.80	0.73	1.5
18 or Older	Total	0.21	-0.44	1.9
	Male	0.73	-0.29	2.8
	Female	-0.30	-0.60	1.1

IC = item count.

To further examine the effect of verification on the IC response, the proportion of original responses that were changed and the direction of the change was estimated. ***Table 9.4*** summarizes the results. Overall, about 1 percent of all responses changed in the verification process. Further, twice as many respondents decreased the originally reported count as increased it. Changes from a count of 0 to 1 or 1 to 0 accounted for more than half of the revisions.

Table 9.4 Percentage of Responses Changed by Verification and the Direction of the Change

Item Count Question	Changed in Verification (Percent)	Changes Revised Downward (Percent)	Changes Revised Upward (Percent)
ICQ1(S)	0.93	70.9	29.1
ICQ1(L)	0.69	56.5	43.5
ICQ2(S)	1.11	73.8	26.2
ICQ2(L)	0.74	64.5	35.5
Average	0.87	66.4	33.6

In addition to bias, variance may be another problem contributing to the unexpected results in ***Table 9.3***. This variance is due partly to sample weighting and partly to response variance or unreliability. For example, most of the negative estimates in the table occur for the weighted estimates, suggesting that the unequal weighting effect increased the imprecision of the estimates.

In ***Table 9.5***, the SEs of weighted traditional estimates are presented for the IC estimates and the survey estimates. The SEs in the table are considerably inflated from their theoretical expectations, presumably as a result of measurement variance. It can be shown that if the reliability of a measure is R, the variance of the estimate of the mean based upon that measure is increased by approximately the factor $1/R$. As an example, suppose the reliability of the IC measures is approximately 0.5. Then the variance of the IC estimate of cocaine use is increased by a factor of about 2—a 100 percent increase in variance. This is equivalent to a 50 percent reduction in sample size for the experiment. Therefore, unreliability can be a major contributor to the instability of an estimator. Response inconsistency may be an important reason for the negative estimates in ***Tables 9.2*** and ***9.3***.

To further examine this point, the reliability of both IC questions was computed. The traditional method for assessing the reliability of a question is through a test-retest design. A form of test-retest data is available from the IC experiment because the items making up the IC short-list questions also were asked individually for the same respondents. The responses to the individual items can be used to form a second count of the IC short-list items. For example, let y_k for $k = 1,...,5$ denote a response to question k corresponding to item k in one of the IC short-list questions, where $y_k = 0$ if the response is "no" and 1 if the response is "yes." Then, if there is no nonsampling error in either the response to the IC short-list question or the corresponding individual questions, y_k,

Table 9.5 Standard Errors of the Weighted Estimates, by Age and Gender

Age	Gender	Weighted IC	2001 Survey
12 to 17	Total	0.49	0.10
	Male	0.75	0.14
	Female	0.63	0.15
18 or Older	Total	0.39	0.09
	Male	0.63	0.15
	Female	0.45	0.08

IC = item count.

then $\Sigma_k y_k$ should be equal to the response to the IC short-list question. Thus, the reliability of the IC short-list question count can be evaluated from the individual question data.

Treating the IC and pseudo-IC responses as test-retest data, the reliability of the ICQ1(S) and ICQ2(S) questions was computed using the Cohen's kappa. For ICQ1(S), kappa is 0.48, and for ICQ2(S) it is 0.43, which are both quite low. This suggests that measurement error variance is a serious problem with the IC approach and may be an important contributor to the failure of the approach to produce valid estimates of cocaine use. Using the reliability inflation factor $1/R$ above suggests that the variance inflation due to measurement error may be 2 or more—equivalent to a 40 percent SE increase.

Table 9.6 compares the IC response to the pseudo-IC response (i.e., the values of $\Sigma_k y_k$, for both short IC questions). The table indicates a considerable amount of inconsistency between the two responses as only 84.9 percent of the responses are in agreement. Among the disagreements (shown in the cells that are off the diagonal of the table), the pseudo-IC response is higher than the IC response for approximately 75 percent of the cases. Many of the differences are quite extreme. For example, 1,393 persons responded "no" (0) to the IC question but answered "yes" (1) to all four items when the questions were asked individually. This large inconsistency is even more puzzling when compared with the number of persons who responded "yes" to two or three individual items—718 and 263, respectively. If the primary cause of the inconsistencies is memory error, the number of persons who recorded "0" to the IC question and then answered with three positives to the individual questions should be greater than the number of persons who answered with four positives to the individual questions. This suggests that 1,393 individuals in the 0-4 cell of the table may have been confused as to how to respond to the IC question.

Table 9.6 Item Count Response, by Pseudo-Item Count Response for Both Short IC Questions

Pseudo-IC Response	Item Count Response				
	0	1	2	3	4
0	51,015	1,641	286	49	47
1	4,392	6,333	447	48	19
2	718	607	622	53	9
3	263	114	48	44	7
4	1,393	96	37	9	8

IC = item count.

A Possible Approach for Compensating for Measurement Error

These preliminary results suggest that the failure of the IC methodology to reduce the bias in estimates of cocaine use is due to measurement error in the IC responses. If the response distribution to the IC questions could be adjusted to compensate for such errors, then less biased and more stable estimates of cocaine use could be obtained. One possibility for incorporating a classification error adjustment into the IC estimation process is through the use of latent class models (e.g., see Heinen, 1996; Vermunt, 1997).

It can be shown that the standard IC estimator is a function of the response distribution to a short-list question, ICQ1(S) or ICQ2(S), and a long-list question, ICQ1(L) or ICQ2(L). Latent Class Analysis (LCA) can be used to correct the response distributions to these questions for measurement error by incorporating in the model the IC question responses, the pseudo-IC

responses, and the responses to the direct cocaine question. Essentially, the model uses the information from these responses to estimate the classification error associated with the direct cocaine question response. More importantly, the model yields a measurement-error-corrected estimate of cocaine use prevalence.

To briefly describe the general idea, let *X* denote the true short-list response and let *Z* denote the true long-list response to one of the IC lists of items. Instead of observing *X* and *Z* in the survey, *A* and *D*, the reported IC response to the short- and long-list questions, are observed, respectively. As is apparent from ***Table 9.2***, IC estimates based upon *A* and *D* are biased as a result of measurement error. Because *X* and *Z* are assumed to be error free, IC estimates based upon *X* and *Z* will be unbiased.

Thus, the goal of the modeling approach is to correct the distribution of responses *A* and *D* for measurement error to estimate the distributions of *X* and *Z*. To obtain an identifiable model (i.e., a model that produces unique estimates of the parameters), a second indicator of *X* is needed. A second indicator is provided by the pseudo-item count variable described above, which is denoted *B*. The NSDUH response to the past year cocaine use question also must be employed in the LCA model in order to obtain an identifiable model. Denote the cocaine response *C* and denote the true status of past year cocaine use by *Y* (i.e., *C* is an indicator of *Y*).

One complication in this analysis is that *A* and *B* are obtained for half of the sample and *D* is obtained for the other half. Because *C* is obtained for the entire sample, this variable can be used to link the two half-samples together to produce an identifiable model. This method will produce a maximum likelihood estimate of $P(Y = 1)$, the true proportion of the population who used cocaine in the past year.

Future research on the latent IC method for NSDUH will focus on model selection and identifiability, estimation of SEs of the estimates, and the robustness of the estimates when the model assumptions are not fully satisfied.

Conclusions

Considerable effort was directed toward the development, implementation, and analysis of an IC methodology for the estimation of cocaine use prevalence in NSDUH. Several adaptations of existing methods were implemented, offering hope that the refined method would succeed in improving the accuracy of prevalence estimation. These adaptations included taking advantage of the ACASI administrative mode to allow respondents to perform a practice item before proceeding to the actual items and asking respondents for confirmation of their answers.

One of the greatest challenges of the study was developing lists of behaviors for the IC that are seen as appropriately contextual with cocaine use and are either very high or very low in prevalence, but not to the point of being perceived as outrageous or implausible. The size of the lists of risky behaviors was carefully optimized, and each item within the lists was cognitively tested. The task was set up in such a way as to eliminate the need for a complex explanation to respondents, thus eliminating potential confusion and mistrust. This was done by including the cocaine use item in a set of contextually relevant items. The target item, regarding use of cocaine, was carefully crafted so that the potential for misinterpretation was minimized. After

calculating the required sample size, the data were collected and analysis was performed, comparing the standard calculation of estimated prevalence and the standard IC method.

Despite these efforts, the IC methodology failed to produce estimates of cocaine use that were even at the level of those obtained by simply asking respondents directly about their cocaine use. Because the direct questioning method is believed to produce underestimates of cocaine, these findings suggest that the IC methodology is even more biased than self-reports.

Because each item in the short-list IC question was asked separately for each respondent, it was possible to check the response consistency of the short-list IC question and a pseudo-IC response formed as a count of positive responses to the four individual questions. This comparison showed that the IC questions were answered quite unreliably, which would explain the failure of the IC approach to yield satisfactory results. This may have been a result of respondents' failing to give careful thought to each item in the IC list when counting the number of applicable behaviors. Then, too, it is possible that IC question responses were affected by context effects (i.e., when an item is considered in a separate question, its interpretation by respondents may be different from when it is part of a list of items thought to be similar on some dimension). It also is possible that substance users were more suspicious about the task than originally envisioned.

Although there is no way to determine all of the reasons that the IC results were unsatisfactory, it is clear that measurement error played a role. Given this situation, an obvious question is whether it is possible to correct the IC estimates for measurement error bias. LCA can be used to estimate and adjust for the measurement error in the IC response process. To that end, an LCA model was proposed that incorporated information from the short and long IC questions, the direct questions for each short-list item (in the form of a pseudo-IC response), and the responses to the direct cocaine question. Preliminary results suggest that this approach has some merit. However, more research and analysis is needed to better model the interactions among these items and better understand the properties of the resulting estimates.

References

Biemer, P. P., & Wiesen, C. (2002). Measurement error evaluation of self-reported drug use: A latent class analysis of the US National Household Survey on Drug Abuse. *Journal of the Royal Statistical Society, Series A (Statistics in Society), 165*(1), 97-119.

Clark, D. C., Sommerfeldt, L., Schwarz, M., Hedeker, D., & Watel, L. (1990). Physical recklessness in adolescence: Trait or byproduct of depressive/suicidal states? *Journal of Nervous and Mental Disease, 178*, 423-433.

Droitcour, J., Caspar, R. A., Hubbard, M. L., Parsley, T. L., Visscher, W., & Ezzati, T. M. (1991). The item count technique as a method of indirect questioning: A review of its development and a case study application. In P. P. Biemer, R. H. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 185-210). New York: Wiley.

Gonzalez, J., Field, T., Yando, R., Gonzalez, K., Lasko, D., & Bendell, D. (1994). Adolescents' perceptions of their risk-taking behavior. *Adolescence, 29*, 701-709.

Greenberg, B. G., Abul-Ela, A. A., Simmons, W. R., & Horvitz, D. G. (1969). The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association, 64*, 520-539.

Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences* (Vol. # 6, Advanced Quantitative Technology in the Social Sciences Series). Thousand Oaks, CA: Sage Publications.

Lewinsohn, P. M., Langhinrichsen-Rohling, J., Langford, R., Rohde, P., Seeley, J. R., & Chapman, J. (1995). The life attitudes schedule: A scale to assess adolescent life-enhancing and life-threatening behaviors. *Suicide & Life-Threatening Behavior, 25*, 458-474.

Shimizu, I. M., & Bonham, G. S. (1978). Randomized response technique in a national survey. *Journal of the American Statistical Association, 73*(361), 5-39.

Vermunt, J. K. (1997). *Log-linear models for event histories* (Vol. # 8, Advanced Quantitative Technology in the Social Sciences Series). Thousand Oaks, CA: Sage Publications.

Wiseman, F., Moriaty, M., & Schafer, M. (1975-1976). Estimating public opinion with the randomized response model. *Public Opinion Quarterly, 39*, 507-513.

10. Comparing NSDUH Income Data with Income Data in Other Datasets

Alexander J. Cowell and Daniel Mamo[1]
RTI International

Introduction

Personal income and family or household income are among the many demographic measures obtained in the National Survey on Drug Use and Health (NSDUH).[2] Because income is a correlate of substance use and other behaviors, it is important to evaluate the accuracy of the income measure in NSDUH. One metric of accuracy is to compare the estimated distribution of income based on NSDUH with the distributions from other data sources that are used frequently. This chapter compares the distribution of 1999 personal income data from the 2000 NSDUH with the distributions in the same year from the Current Population Survey (CPS) and the Statistics of Income (SOI) data.

NSDUH is conducted annually by the Substance Abuse and Mental Health Services Administration (SAMHSA), Office of Applied Studies (OAS), and is the primary source of statistical information on substance use and abuse by the U.S. population. It collects information through face-to-face interviews from residents of households, noninstitutional group quarters (e.g., shelters, rooming houses, dormitories), and civilians living on military bases. Persons excluded from the survey include homeless persons who do not use shelters, military personnel on active duty, and residents of institutional group quarters, such as prisons and long-term hospitals. Data are collected according to a stratified, multistage area probability design. NSDUH's sample is representative of almost 98 percent of the U.S. population aged 12 years old or older. The 2000 survey had an overall response rate of 73.93 percent and an item response rate for the income measure used in this analysis of 93.47 percent among survey respondents. Missing values were imputed using the predictive mean neighborhood (PMN) method, a combination of the non-model-based hot deck and a modification of the modal-assisted predictive mean matching method. Analysis weights allow for the production of national estimates at the person and household ("dwelling unit") levels. The 2000 survey collected information from 71,764 persons. For more information, see OAS (2001).

The CPS is a nationwide survey of households conducted monthly for the U.S. Bureau of Labor Statistics (BLS) by the U.S. Bureau of the Census. The CPS core survey is the primary source of information on key employment characteristics in the United States, including unemployment estimates released every month by the BLS. An annual supplement to the March

[1] Now with United Guaranty Corporation.
[2] Prior to 2002, the survey was called the National Household Survey on Drug Abuse (NHSDA). In this report, it is referred to as NSDUH, regardless of the survey year.

survey collects income and other demographic information and is included on the CPS's Annual Demographic Survey File. The Annual Demographic Survey File is used in this chapter's analysis.

The main CPS has a universe very similar to NSDUH: the civilian, noninstitutionalized population of the United States. Both surveys, for example, include shelters in their definition of group quarters. However, there are some differences. Both the main CPS and its Annual Demographic Survey gather information on persons 15 years or older compared with 12 years or older in NSDUH. The Annual Demographic Survey also includes a sample of active-duty members of the Armed Forces who live off base, as well as those who live on base with their families.

There also are some differences between NSDUH and the CPS in their survey designs and methods of data collection. Although the CPS also employs a stratified, multistage, probability design, it surveys households, not individuals. Information is obtained on all household members, usually via a single respondent. And although most information is gathered during a face-to-face interview, about 15 percent of CPS respondents are surveyed over the telephone. Analysis weights allow for estimation at the person, family, and household levels.

The 2000 CPS Annual Demographic Survey had a response rate of 85.6 percent and collected information on 133,710 persons. The item response rates for the income variables used to create total personal income ranged from 45.5 percent for interest and dividend income to 72.8 percent for total wages and salary earnings. These items were imputed using a single-item hot-deck method. For more information on the main CPS and the CPS Annual Demographic File, visit the CPS website at http://www.bls.census.gov/cps/cpsmain.htm (BLS & U.S. Bureau of the Census, 2001).

The SOI data do not come from a survey, but are based on an annual stratified national sample of all individual income tax returns (Forms 1040A, 1040EZ, and 1040PC) filed by U.S. citizens and residents. The Revenue Act of 1916 required the U.S. Internal Revenue Service (IRS) to publish annual statistics, a requirement that remains in effect until this day (Internal Revenue Code, Section 6108). The IRS meets this requirement by publishing tabulations of the SOI data (IRS, 2003a). The main income measure available in the SOI is the adjusted gross income (AGI) of the tax-filing unit. The 2000 sample contained information on 191,010 tax-filing units and captured about 97.3 percent of all tax returns, all of which reported income. For more information on the 2000 SOI data, see the IRS website (IRS, 2003a).

NSDUH and the SOI differ in three important ways that affect the analysis presented here: the unit of analysis, the likelihood that income is captured by the data, and the filing threshold. First, NSDUH represents individuals, whereas the SOI represents tax-filing units. Whether a tax-filing unit is the same as an individual depends on the tax-filing status, which in turn depends on marital status and family circumstances. Second, NSDUH data are obtained from a self-report survey, where respondents have no legal obligation to report truthfully; although the SOI data also are self-reported, respondents are legally accountable for their responses. Third, and related to the second difference, people earning below certain income thresholds are not required to file tax returns; thus, the SOI data may underrepresent lower income people. This filing threshold depends on filing status and age and is discussed more fully in the section on methods.

Relevant Literature

The literature provides little specific guidance on comparing income in NSDUH with income in either the CPS or the SOI. Although no research to date compares NSDUH income data with other sources, a body of research exists comparing CPS data with the SOI, as well as with other datasets. This literature informs the task at hand because the measure of personal income in the CPS is very similar to that in NSDUH. This brief literature review delineates the core issues in comparing the CPS with the SOI and highlights the methods that have been used to compare data sources. For the literature review, electronic databases and the Internet were searched for literature comparing the CPS or other survey measures of income with the SOI.

Just as when comparing NSDUH with the SOI, one primary concern in comparing the CPS with the SOI is that the intended purpose of income reported in the two datasets is fundamentally different (Irwin & Herriot, 1982). Income reported in the CPS is used for research purposes, whereas the AGI reported in the SOI is used to determine taxes. Given the difference in purpose, one should expect differences in the distributions. Because the purpose of the reported data is different, the unit of analysis in the SOI differs from the CPS and NSDUH. The unit of analysis in the SOI is a tax-filing unit, which is sometimes two people (as with tax returns filed jointly by married couples). By comparison, the CPS provides income data at the individual, family, and household levels (Mann & Salvo, 1984).

One important finding from the literature is that direct comparisons of income distributions in survey data and in the SOI are likely to be less accurate for people in either the lower income or higher income intervals (Childers & Hogan, 1984; David & Triest, 1983; Irwin & Herriot, 1982). Lower income groups may not provide a good comparison because many people in these groups also are young and single, a relatively mobile demographic group. Because of their mobility, lower income filers often are underrepresented in surveys (Childers & Hogan, 1984). However, regardless of their mobility, people who are young and in the lower income group still have to file tax returns. Therefore, because such filers may be underrepresented in surveys, the income distribution for this group as reported from surveys, such as NSDUH and the CPS, is unlikely to match that shown in the SOI. Another possible reason for a potential disparity in lower income intervals that is not addressed in the literature is that people with incomes below certain thresholds are not required to file returns.

With regard to higher income intervals, David and Triest (1983) demonstrated that the distribution in the CPS and the SOI is likely to be very different. Although the authors found some evidence of a linear relationship at lower income intervals, they found strong nonlinearities in the relationship between CPS and SOI data for higher income intervals. The authors determined that the main reason for this nonlinearity is that underreporting of income increases as income levels rise. This finding suggests that the income distribution in NSDUH is unlikely to match that in the SOI for the higher income intervals. The likely mismatch depends on the breadth of the intervals over which the two distributions are compared, as discussed in the section on methods.

Although a number of statistical techniques are available for comparing income distributions, the methods favored by researchers comparing the CPS to the SOI are straightforward. All the research specifically comparing these two datasets relied on ordinary least squares regression,

tabulations, and means comparisons (Childers & Hogan, 1984; David & Triest, 1983; Irwin & Herriot, 1982; Mann & Salvo, 1984; Park, 2002).

Methods

Table 10.1 summarizes the income measures used in the three datasets. The income measure used from the two surveys (NSDUH and the CPS) is personal income, rather than family or household income. Although the CPS explicitly gathers information about more sources of personal income than NSDUH, the income sources in the CPS were combined to map exactly to NSDUH. The income measure from the SOI is AGI, reported by various combinations of filing status.

Table 10.1 Measure of Income and Reported Income Intervals for NSDUH, CPS, and SOI Single Filers

	NSDUH		CPS	SOI
Measure of Income	In the 2000 survey, total personal income is equal to income from the 1999 calendar year. Income is defined as earned wages, social security, supplemental security income, public assistance, savings of dividend income, child support, and other sources.		In the March 2000 CPS, total personal income is equal to income from April 1999 through March 2000. Income is defined as wages and salaries, unemployment, worker's compensation, social security, welfare, veterans' benefits, disability, retirement, interest and dividends, rental income, education income, child support, and alimony.	AGI of single filers for the 1999 calendar year.
Reported Income Intervals	Less than $1,000 $1,000–$1,999 $2,000–$2,999 $3,000–$3,999 $4,000–$4,999 $5,000–$5,999 $6,000–$6,999 $7,000–$7,999 $8,000–$8,999 $9,000–$9,999 $10,000–$10,999 $11,000–$11,999 $12,000–$12,999 $13,000–$13,999 $14,000–$14,999	$15,000–$15,999 $16,000–$16,999 $17,000–$17,999 $18,000–$18,999 $19,000–$19,999 $20,000–$24,999 $25,000–$29,999 $30,000–$34,999 $35,000–$39,999 $40,000–$44,999 $45,000–$49,999 $50,000–$74,999 $75,000 or more	The personal income variable is a continuous variable. CPS personal income was coded to fit the income category of the comparison dataset.	Categories below $20,000 depend on filing status reported. Single only: Less than $5,000 $5,000–$9,999 $10,000–$14,999 $15,000–$19,999 All filing statuses: As for NSDUH $20,000–$24,999 $25,000–$29,999 $30,000–$39,999 $40,000–$49,999 $50,000–$74,999 $75,000–$99,999 $100,000–$199,999 $200,000–$499,999 $500,000–$999,999 $1,000,000 or more

AGI = adjusted gross income; CPS = Current Population Survey; SOI = Statistics of Income.

Table 10.1 also summarizes the manner in which income is reported in each data source. The CPS reports income as a continuous measure, whereas NSDUH and the SOI report income in intervals. The SOI intervals are the same as or broader than the NSDUH intervals. For example, in the $30,000 to $50,000 range, the SOI has two intervals and NSDUH has four. Thus,

NSDUH's intervals were aggregated to match those of the SOI. The CPS data also were aggregated to match the SOI intervals.

The three datasets also differ in how higher income intervals are categorized and how income is top-coded. The highest income interval reported in NSDUH is $75,000 or more. The CPS and SOI record higher income amounts, but these were collapsed to be the same as the highest income interval in NSDUH. This approach to top-coding the data will help mitigate the problem noted by David and Triest (1983) that income tends to be underreported at higher income levels.

The IRS (2003b) defines five filing statuses, which depend on marital and family circumstances: single, married filing jointly, married filing separately, head of household, and qualifying widow(er). The SOI data are reported by certain filing statuses. Two SOI filing statuses were used in this chapter: single filers only and filers of all statuses.

The income intervals used depend on the filing status for which the SOI data are reported. When the SOI data are reported regardless of filing status, income up to $20,000 is reported in $1,000 increments, matching the NSDUH intervals. However, when the SOI data are reported for single filers only, income in this range is reported in $5,000 increments.

As noted in the Introduction, people earning below certain income thresholds are not required to file tax returns. The SOI data may therefore underrepresent lower income people. ***Table 10.2*** summarizes the filing thresholds by age and filing status. For example, in 1999, a person whose filing status was single and who was under age 65 was only required to file a tax return if his or her income was greater than or equal to $7,050. Thus, for single filers, the number of observations in the two lowest income intervals in the SOI is expected to be lower than that in either of the surveys.

Table 10.2 Minimum Income Requirements for Filing a 1999 Tax Return

If filing status is...	...and age is...	...then a return must be filed if gross income was at least
Single	Younger than 65	$7,050
	65 or older	$8,100
Married Filing Jointly	Both spouses younger than 65	$12,700
	One spouse 65 or older	$13,550
	Both spouses 65 or older	$14,400
Married Filing Separately	All ages	$2,750
Head of Household	Younger than 65	$9,100
	65 or older	$10,150
Qualifying Widow(er)	Younger than 65	$9,950
	65 or older	$10,800

Two sets of comparisons were made that differed by marital status in the surveys and filing status in the SOI. First, the income distribution reported in the surveys, regardless of marital status, was compared with that in the SOI, regardless of filing status. For those tax-filing units whose status is "married filing jointly," the reported AGI will be for two people, whereas the survey data used are for individual income only. Consequently, these comparisons are unlikely to provide a close fit of the income distributions, particularly in higher income intervals. Second, the income distribution in the surveys for those who are unmarried (excluding widow[er]s) was compared with the income distribution in the SOI for those whose filing status is single. Because only unmarried people can file as "single," the restrictions in this second set of comparisons should ensure a relatively close fit between the survey data and the SOI. The data do not allow a

reasonable comparison to be made between the income distribution in NSDUH and the distribution for other filing statuses in the SOI, such as "married, filing jointly." Because pair-level weights were not available for NSDUH at the time of this analysis, a reporting unit in NSDUH could not be created so that it could compare reasonably to "married filing jointly" in the SOI. NSDUH weights are calibrated to represent individuals. For NSDUH data to represent a pair of married people, rather than individuals, a different set of weights—pair-level weights—are required.

In both sets of comparisons, those under age 16 were excluded from the survey data to make as close a comparison of income distribution as possible. People under age 16 rarely file income returns (David Jordan, Statistical Information Services Office, IRS, personal communication with Daniel Mamo, RTI, 2003) probably because most of those younger than 16 who earn income do not earn enough to meet the thresholds for filing (Emily Gross, Statistical Information Services Office, IRS, personal communication with Daniel Mamo, RTI, 2003).

The primary method of comparing the income distributions for this preliminary analysis was to examine histograms of the weighted income distributions from the three datasets. Two-way and three-way comparisons were made. Comparisons are based only on visual inspection of the data, without formal testing for statistical significance or calculation of sampling error.

Results

Figures 10.1 through ***10.4*** present the first set of comparisons between the datasets (for all marital statuses and all filers). ***Table 10.3*** contains the number of observations and the frequencies underlying these figures. ***Figures 10.5*** through ***10.8*** present the second set of comparisons.

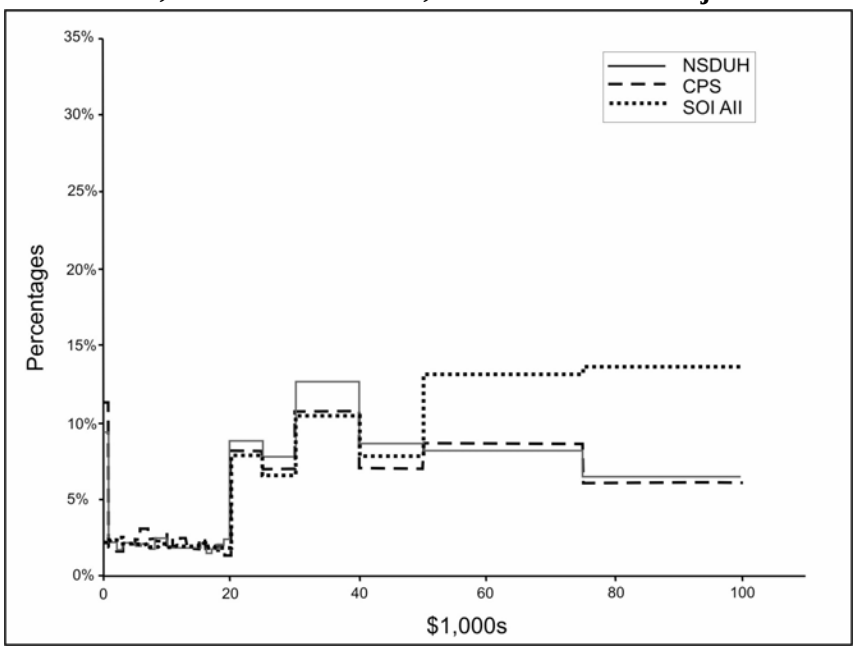
Figure 10.1 presents histograms of the distribution from all three datasets. It suggests two general findings. First, as expected, the income distributions in the two surveys (NSDUH and CPS) are very close to one another. Second, the survey distributions differ considerably from the SOI in the lowest income interval and in the two highest income intervals.

For the lowest income interval (less than $1,000), there are large differences between the estimates from the surveys and the SOI. This interval contains 9.4 and 11.4 percent of the NSDUH and CPS samples, respectively. By comparison, only 2.3 percent of the SOI sample is in this income interval.

For income levels between $1,000 and $29,999, all three distributions appear close to one another. The $30,000 to $49,999 range contains a higher percentage of the NSDUH sample than either the CPS or SOI. In the two highest income intervals, another large difference is seen between the percentage in the surveys and the percentage in the SOI. The percentage of the sample in the $50,000 to $74,999 range is between 8 and 9 percent for the surveys, whereas it is just over 13 percent for the SOI. Similar frequencies are seen for the over $75,000 income interval.

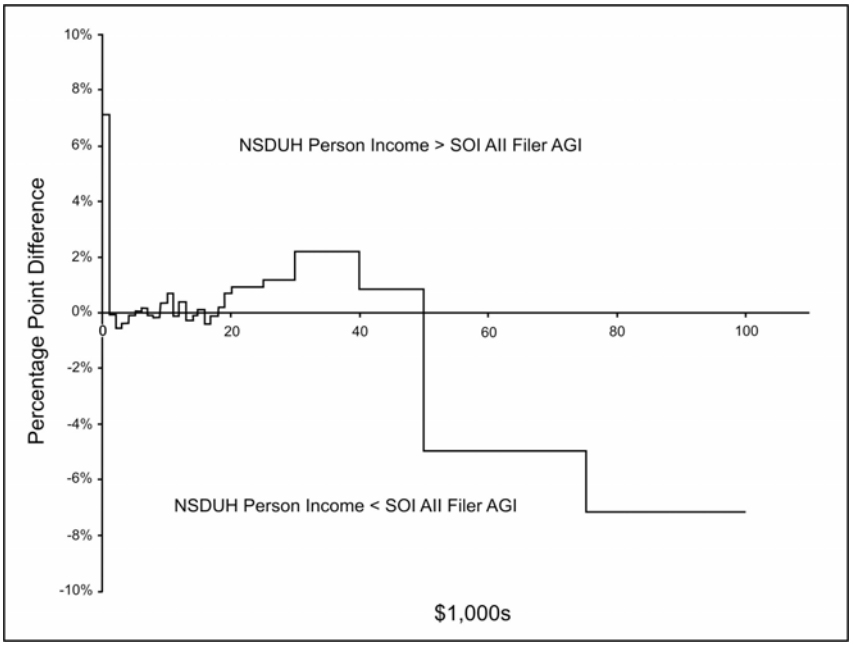
Figures 10.2 through ***10.4*** present the percentage point differences when making the pairwise comparisons between the three income series shown in ***Figure 10.1***. The figures are

Figure 10.1 Comparison of Income Distribution for Individuals 16 Years or Older (NSDUH Person Income, CPS Person Income, and SOI All Filer Adjusted Gross Income)



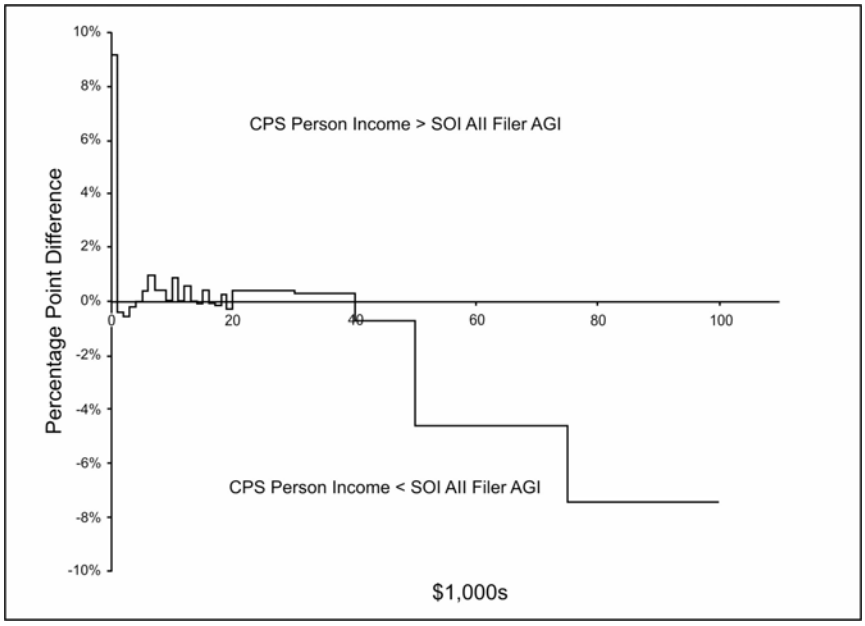
Data Sources: 2000 National Survey on Drug Use and Health (NSDUH) produced by SAMHSA, March 2000 Current Population Survey (CPS) produced by the Bureau of Labor Statistics and the U.S. Bureau of the Census, and the 1999 Statistics of Income (SOI) produced by the U.S. Internal Revenue Service.

Figure 10.2 Percentage Point Difference in the Comparison of Income Distribution for Individuals 16 Years or Older (NSDUH Person Income–SOI All Filer Adjusted Gross Income)



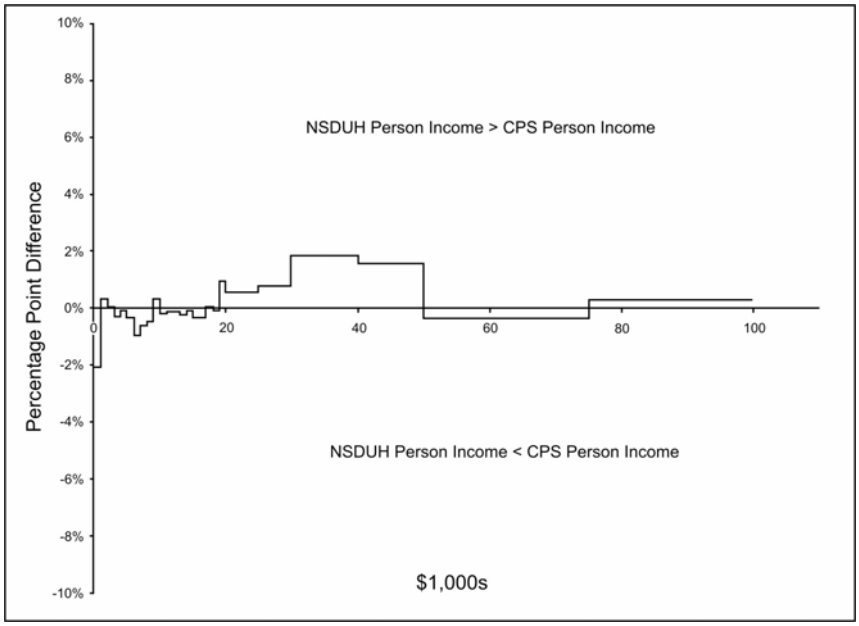
Data Sources: 2000 National Survey on Drug Use and Health (NSDUH) produced by SAMHSA and the 1999 Statistics of Income (SOI) produced by the U.S. Internal Revenue Service.

Figure 10.3 Percentage Point Difference in the Comparison of Income Distribution for Individuals 16 Years or Older (CPS Person Income–SOI All Filer Adjusted Gross Income)



Data Sources: March 2000 Current Population Survey (CPS) produced by the Bureau of Labor Statistics and the U.S. Bureau of the Census and the 1999 Statistics of Income (SOI) produced by the U.S. Internal Revenue Service.

Figure 10.4 Percentage Point Difference in the Comparison of Income Distribution for Individuals 16 Years or Older (NSDUH Person Income–CPS Person Income)



Data Sources: 2000 National Survey on Drug Use and Health (NSDUH) produced by SAMHSA and March 2000 Current Population Survey (CPS) produced by the U.S. Bureau of Labor Statistics and the U.S. Bureau of the Census.

Table 10.3 Weighted Comparison of Income Distribution for All Individuals 16 Years Older in NSDUH and the CPS Compared with All Filing Statuses in the SOI

	All NSDUH		All CPS		All SOI Filers	
Income Category	Weighted *N*	Percentage Distribution	Weighted *N*	Percentage Distribution	Weighted *N*	Percentage Distribution
Under $1,000	19,430,068	9.37%	23,955,872	11.42%	2,880,330	2.27%
$1,000–$1,999	4,561,200	2.20%	3,946,785	1.88%	2,909,501	2.29%
$2,000–$2,999	3,822,139	1.84%	3,792,390	1.81%	3,012,426	2.37%
$3,000–$3,999	3,716,339	1.79%	4,321,233	2.06%	2,854,708	2.25%
$4,000–$4,999	4,224,468	2.04%	4,491,670	2.14%	2,759,177	2.17%
$5,000–$5,999	4,235,848	2.12%	5,022,527	2.39%	2,570,135	2.02%
$6,000–$6,999	4,430,779	1.80%	6,487,725	3.09%	2,650,302	2.09%
$7,000–$7,999	3,742,842	1.92%	5,113,078	2.44%	2,539,115	2.00%
$8,000–$8,999	3,988,450	2.43%	5,068,570	2.42%	2,562,949	2.02%
$9,000–$9,999	5,049,243	2.43%	4,447,652	2.12%	2,657,214	2.09%
$10,000–$10,999	5,337,597	2.57%	5,790,004	2.76%	2,411,630	1.90%
$11,000–$11,999	3,722,235	1.79%	4,018,235	1.91%	2,471,051	1.94%
$12,000–$12,999	4,921,391	2.37%	5,227,970	2.49%	2,486,017	1.96%
$13,000–$13,999	3,607,786	1.74%	4,102,140	1.95%	2,466,393	1.94%
$14,000–$14,999	3,665,160	1.77%	3,879,301	1.85%	2,440,627	1.92%
$15,000–$15,999	4,437,434	2.14%	5,163,611	2.46%	2,588,996	2.04%
$16,000–$16,999	3,172,590	1.53%	3,919,033	1.87%	2,433,853	1.92%
$17,000–$17,999	3,713,756	1.79%	3,625,467	1.73%	2,372,806	1.87%
$18,000–$18,999	4,006,340	1.93%	4,235,833	2.02%	2,236,508	1.76%
$19,000–$19,999	4,941,228	2.38%	2,964,085	1.41%	2,151,011	1.69%
$20,000–$24,999	18,211,569	8.78%	17,266,054	8.23%	9,967,211	7.84%
$25,000–$29,999	16,078,442	7.75%	14,693,817	7.00%	8,392,769	6.60%
$30,000–$39,999	26,148,676	12.60%	22,585,053	10.76%	13,288,379	10.46%
$40,000–$49,999	17,874,498	8.62%	14,789,527	7.05%	9,870,199	7.77%
$50,000–$74,999	17,037,492	8.21%	18,000,174	8.58%	16,755,560	13.19%
$75,000 or More	13,421,001	6.47%	12,939,473	6.17%	17,346,280	13.65%

CPS = Current Population Survey; SOI = Statistics of Income.

derived directly from ***Figure 10.1***, but they also are informative in their own right. ***Figure 10.2*** demonstrates the percentage point difference between the frequencies in NSDUH and the SOI. ***Figure 10.2*** shows that in the lowest income interval, the distribution from NSDUH is over 7 percentage points higher than that from the SOI. Similarly, for the two highest income intervals, the distribution from the SOI is over 5 percentage points higher than that reported in NSDUH. Other interval comparisons show far lower differences between the distributions; the percentage point difference for all other intervals is 2 percentage points or lower.

Figure 10.3 presents a similar comparison between the CPS and SOI distributions. A similar pattern to the NSDUH–SOI two-way comparison is seen. The percentage of the CPS sample in the lowest interval is over 9 percentage points higher than the percentage of the SOI in that interval. The two highest intervals account for a higher proportion of the CPS sample than in the SOI by approximately 4 and 8 percentage points, respectively. The absolute value of the difference in the proportion of the samples in the other intervals is very low, at less than 1 percentage point.

Figure 10.4 shows the comparison between NSDUH and the CPS. The distributions are very close; in no interval does the absolute value of the difference exceed 2 percentage points.

Figure 10.5 displays the three income distributions for the second set of comparisons, when the survey samples are restricted to single people and the SOI sample is restricted to single filers. ***Table 10.4*** contains the number of observations and the frequencies underlying ***Figures 10.5*** through ***10.8***. In general, it is evident that restricting the comparison to single people/filers vastly improves the match among the three distributions. With the exception of the lowest income interval (those earning below $5,000), the frequencies of the three datasets are very close to one another across all income intervals. Although not as close as for the other income intervals, the frequencies in the lowest income interval are reasonably close: 28.8 percent in the CPS, 25.2 percent in NSDUH, and 21 percent in the SOI.

The closeness-of-fit between the three distributions also is evident in ***Figures 10.6*** through ***10.8***, which show the two-way comparisons of the income distributions. Apart from the lowest income interval, the absolute value of the difference between any two distributions is no larger than 2.3 percentage points.

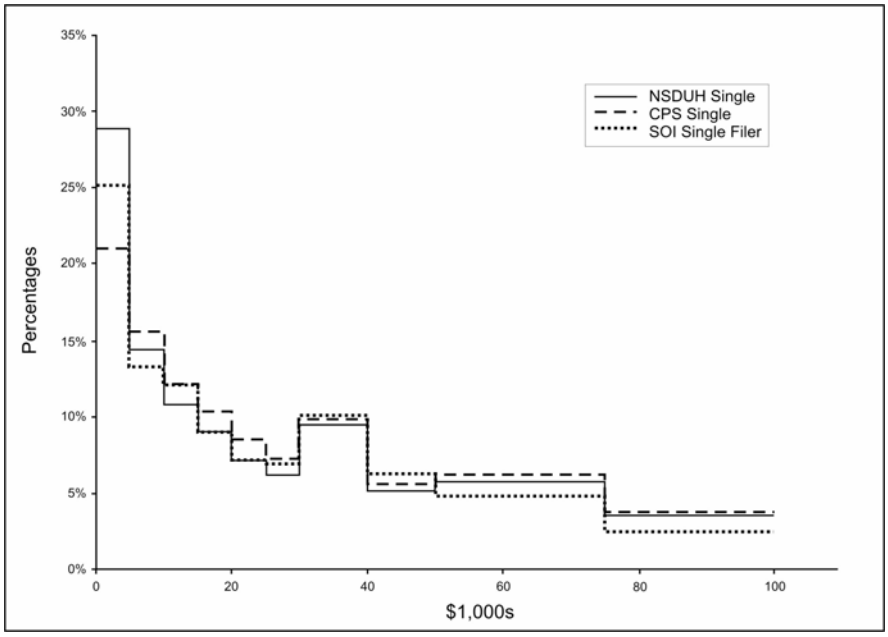
Discussion and Conclusions

Despite some fundamental differences between the SOI and either of the survey datasets (CPS and NSDUH), there are strong similarities between the three income distributions. In both sets of comparisons, the frequencies reported in NSDUH and the CPS are typically within 2 percentage points of each other across all income intervals. With the exception of the lowest interval, in the second set of comparisons (single people and single filers), the frequencies of the three datasets are within 2.5 percentage points of one another across all income intervals.

As expected, it was found that for the second set of comparisons (between single people and single filers) the income distribution from NSDUH compares closely with those from the CPS and SOI. The first set of comparisons (for all marital statuses and all filers) shows some comparatively large disparities between the income distribution in the surveys and that in the SOI. These large disparities are to be expected because the units of analyses are most likely to be different for this set of comparisons. Whereas both NSDUH and the CPS report personal income, the SOI reports income for all tax-filing units, including those whose filing status was "married, filing jointly." Because two incomes combined will be at least as great as any one of them alone, married people filing jointly are more likely to be represented in the higher income intervals. Thus, in the first set of comparisons, there is a relatively large difference between the distributions from the surveys and the distribution from the SOI.

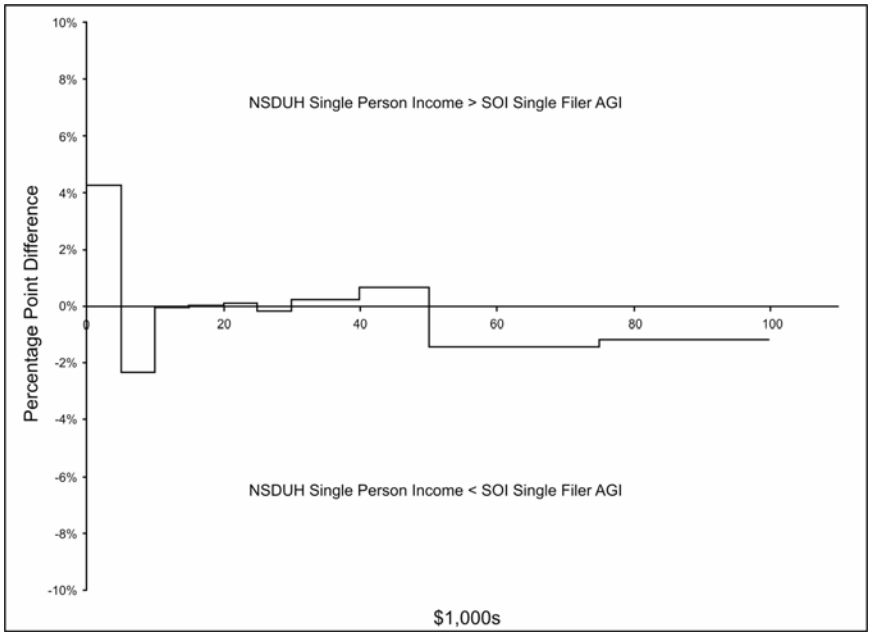
In all comparisons, some of the largest differences in the income distributions were found in the lowest income interval. The literature suggests that the proportion of the sample in this interval would be lower in the surveys than in the SOI. Childers and Hogan (1984), for example, suggested that lower income individuals also tend to be mobile and can be difficult to trace; thus, they may be underrepresented in surveys. Contrary to Childers and Hogan (1984), this chapter's results show that a relatively high proportion of people in the surveys are in the lowest income interval. One reason for this discrepancy may be that lower income individuals are not required to file tax returns. Therefore, lower income individuals are likely to be underrepresented in the SOI compared with NSDUH and the CPS. Another reason for the discrepancy could be that income is generally underreported in surveys. Unlike survey data, income reported to the IRS

Figure 10.5 Comparison of Income Distribution for Unmarried Individuals 16 Years or Older (NSDUH Single Person Income, CPS Single Person Income, and SOI Single Filer Adjusted Gross Income)



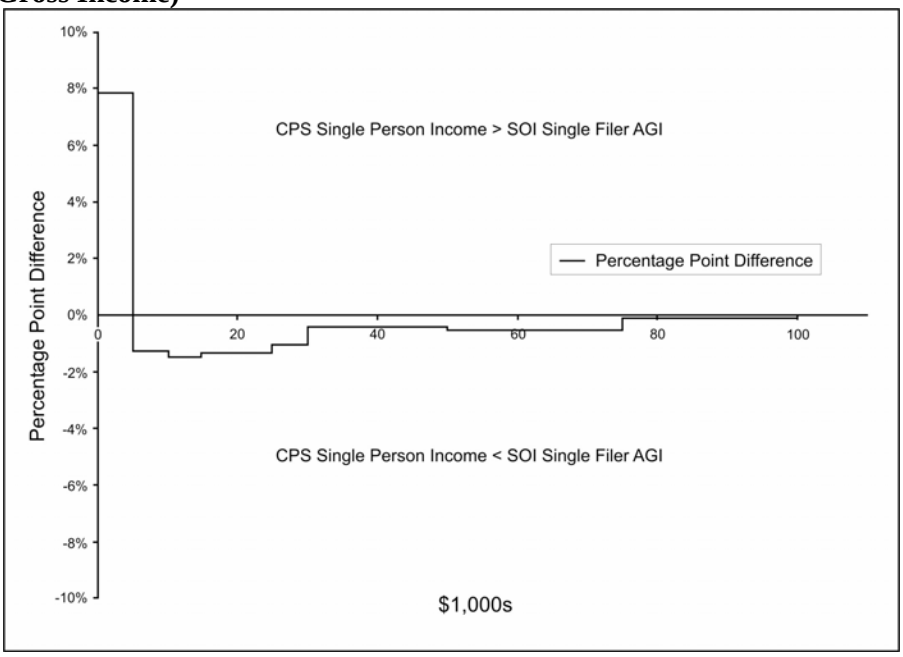
Data Sources: 2000 National Survey on Drug Use and Health (NSDUH) produced by SAMHSA, March 2000 Current Population Survey (CPS) produced by the U.S. Bureau of Labor Statistics and the U.S. Bureau of the Census, and the 1999 Statistics of Income (SOI) produced by the U.S. Internal Revenue Service.

Figure 10.6 Percentage Point Difference in the Comparison of Income Distribution for Unmarried Individuals 16 Years or Older (NSDUH Single Person Income–SOI Single Filer Adjusted Gross Income)



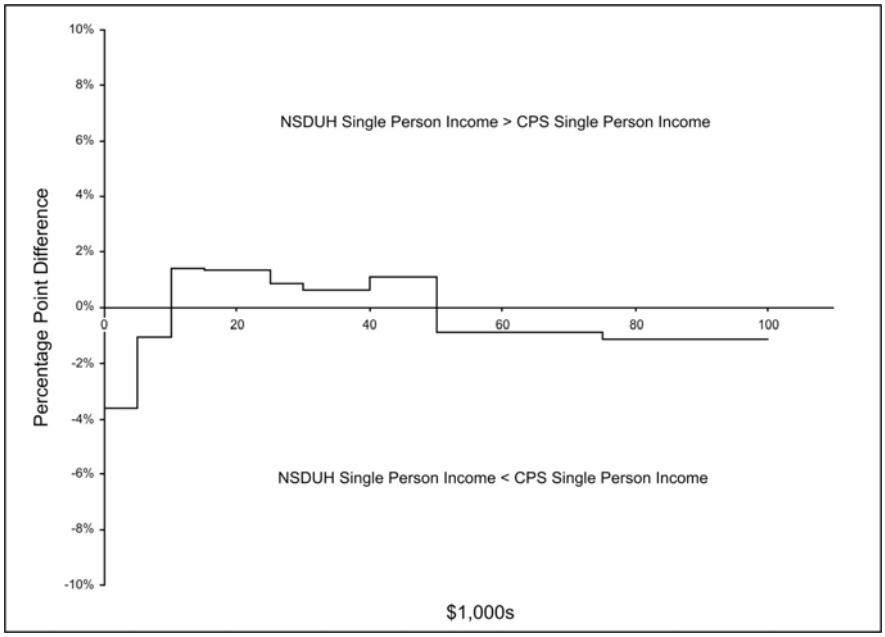
Data Sources: 2000 National Survey on Drug Use and Health (NSDUH) produced by SAMHSA and the 1999 Statistics of Income (SOI) produced by the U.S. Internal Revenue Service.

Figure 10.7 Percentage Point Difference in the Comparison of Income Distribution for Unmarried Individuals 16 Years or Older (CPS Single Person Income–SOI Single Filer Adjusted Gross Income)



Data Sources: March 2000 Current Population Survey (CPS) produced by the U.S. Bureau of Labor Statistics and the U.S. Bureau of the Census and the 1999 Statistics of Income (SOI) produced by the U.S. Internal Revenue Service.

Figure 10.8 Percentage Point Difference in the Comparison of Income Distribution for Unmarried Individuals 16 Years or Older (NSDUH Single Person Income–CPS Single Person Income)



Data Sources: 2000 National Survey on Drug Use and Health (NSDUH) produced by SAMHSA and March 2000 Current Population Survey (CPS) produced by the U.S. Bureau of Labor Statistics and the U.S. Bureau of the Census.

Table 10.4 Weighted Comparison of Income Distribution for Unmarried Individuals 16 Years or Older in NSDUH and the CPS Compared with the Percentage of SOI Single Filers

	Unmarried NSDUH		Unmarried CPS		SOI Single Filers	
Income Category	Weighted *N*	Percentage Distribution	Weighted *N*	Percentage Distribution	Weighted *N*	Percentage Distribution
Under $5,000	19,249,657	25.18%	21,488,937	28.81%	11,928,817	20.95%
$5,000–$9,999	10,160,403	13.29%	10,707,665	14.35%	8,901,541	15.64%
$10,000–$14,999	9,258,085	12.11%	7,988,425	10.71%	6,920,870	12.16%
$15,000–$19,999	7,912,996	10.35%	6,712,401	9.00%	5,878,639	10.33%
$20,000–$24,999	6,493,476	8.49%	5,330,227	7.15%	4,809,595	8.45%
$25,000–$29,999	5,345,296	6.99%	4,583,981	6.14%	4,092,538	7.19%
$30,000–$39,999	7,682,474	10.05%	7,013,374	9.40%	5,595,563	9.83%
$40,000–$49,999	4,778,157	6.25%	3,840,355	5.15%	3,172,118	5.57%
$50,000–$74,999	3,655,938	4.78%	4,243,811	5.69%	3,520,755	6.18%
$75,000 or More	1,903,387	2.49%	2,686,991	3.60%	2,106,710	3.70%

CPS = Current Population Survey; SOI = Statistics of Income.

can be subject to audit; thus, respondents can be held accountable for the accuracy of their reported income.

One direction for future work is to use more sophisticated techniques to compare income distributions beyond the straightforward graphical comparisons presented in this report. An extensive literature on income inequality, for example, uses appropriate techniques that could be applied to a comparison of income distributions (e.g., Ravallion, 1994). Many of these techniques involve examining the fit of the data to a number of alternative parametric distributions that are suitably flexible and general (e.g., Bandourian, McDonald, & Turley, 2002). A second direction for future work is to estimate the sampling error of income distributions and determine the statistical significance in comparisons.

References

Bandourian, R., McDonald, J. B., & Turley, R. S. (2002, June 8). *A comparison of parametric models of income distribution across countries and over time* (Luxembourg Income Study Working Paper 305). Syracuse, NY: Luxemburg Income Study, Maxwell School of Citizenship and Public Affairs. [Available as a PDF at http://www.lisproject.org/publications/wpapersg.htm and http://www.lisproject.org/publications/liswps/305.pdf]

Childers, D. R., & Hogan, H. (1984). Matching IRS records to census records: Some problems and results. In *Proceedings of the American Statistical Association, Survey Research Methods Section* (pp. 301-306). Alexandria, VA: American Statistical Association. [Available as a PDF at http://www.amstat.org/sections/srms/Proceedings/papers/1984_059.pdf]

David, M., & Triest, R. (1983). The CPS hot deck: An evaluation using IRS records. In *Proceedings of the American Statistical Association, Survey Research Methods Section* (pp. 421-426). Alexandria, VA: American Statistical Association. [Available as a PDF at http://www.amstat.org/sections/srms/Proceedings/papers/1983_080.pdf]

Irwin, R., & Herriot, R. (1982). An initial look at preparing local estimates of household size from income tax returns. In *Proceedings of the American Statistical Association, Survey Research Methods Section* (pp. 459-464). Alexandria, VA: American Statistical Association. [Available as a PDF at http://www.amstat.org/sections/srms/Proceedings/papers/1982_088.pdf]

Mann, E. S., & Salvo, J. J. (1984). Linking IRS and census data to produce intercensal income estimates: An exploration. In *Proceedings of the American Statistical Association, Survey Research Methods Section* (pp. 295-300). Alexandria, VA: American Statistical Association. [Available as a PDF at http://www.amstat.org/sections/srms/Proceedings/papers/1984_058.pdf]

Office of Applied Studies. (2001). *Summary of findings from the 2000 National Household Survey on Drug Abuse* (DHHS Publication No. SMA 01-3549, NHSDA Series H-13). Rockville, MD: Substance Abuse and Mental Health Services Administration. [Available at http://www.oas.samhsa.gov/p0000016.htm#standard]

Park, T. S. (2002). *Comparison of BEA estimates of personal income and IRS estimates of adjusted gross income: New estimates for 2000, revised estimates for 1999*. Retrieved July 12, 2004, from http://www.bea.gov/bea/ARTICLES/2002/11November/1102irs&agi.pdf

Ravallion, M. (1994). *Poverty comparisons* (Fundamentals of Pure and Applied Economics Series, Volume 56). Chur, Switzerland: Harwood Academic Publishers.

U.S. Bureau of Labor Statistics & U.S. Bureau of the Census. (2001, April 4). *Home page of Current Population Survey*. Retrieved July 13, 2004, from http://www.bls.census.gov/cps/cpsmain.htm

U.S. Internal Revenue Service. (2003a). *Individual income tax returns, tax year 2000*. Retrieved July 12, 2004, from http://www.irs.gov/taxstats/article/0,,id=96586,00.html

U.S. Internal Revenue Service. (2003b). *IRS Form 501: Exemptions, standard deduction, and filing information for use in preparing 2003 returns*. Retrieved July 12, 2004, from http://www.irs.gov/pub/irs-pdf/p501.pdf