

Question#1: How will bounding be dealt with in a companion collection, when there is no bounding event?

Response: Bounding is the preferable method to control from external telescoping (Neter and Waksberg, 1964). It refers to a procedure in which a prior interview is used to mark the earliest point of the period for which the respondent is expected to report. By referring to the prior interview, the interviewer provides a temporal marker that helps the respondent to define the recounting period and to avoid referencing events that happened before the period of interest. It also allows interviewers to check whether specific incidents are the same as what was reported at the last interview (See Biderman and Cantor 1984.)

Bounding requires interviewing respondents multiple times. In fact, the NCVS is the only victimization survey that incorporates a bounding interview in its design. The British Crime Survey (BCS) has a design that is close to the NCVS in that it uses both a screener and detailed incident form to collect its primary crime estimates. However, it uses a 12-month reference period, with no bound.

Several measures are being utilized to minimize external telescoping on the NSHS:

- ACASI respondents will identify temporal markers by filling out an event history calendar (e.g., Belli, et al, 2001) on paper prior to the sexual victimization screener. Respondents will be asked to consider birthdays, vacations, changes in relationships and jobs, and other significant events that occurred throughout the 12 months prior to the interview, with a particular focus on any events that may have occurred during the anchor month 12 months ago. The goal is to have the respondent think very carefully about the reference period prior to being asked the victimization screener questions. Respondents will then be reminded that they can refer to the calendar as they answer questions about sexual victimization in the past 12 months. This is a process that is similar to that used on the British Crime Survey.
- CATI respondents will not complete the event history calendar, due to the complexity of administering it by phone. However, both the CATI and ACASI instruments have several checks built into them to help the respondent focus just on incidents since the anchor month.
 - Screening items instruct respondents to think about what has happened since an anchor date (e.g. “since February 2013”), rather than asking more generally about the “past 12 months”.
 - If the respondent reports an incident in the past 12 months, they are asked to indicate in which month the most recent 4 incidents occurred. If the respondent indicates it was before the anchor date, the incident is not counted as a past 12 month incident. By asking respondents to date the event, she will be forced to think about the reference period and whether or not it occurred in the past 12 months. This is different from other intimate partner violence surveys which do not confirm the specific dates. Asking for the month of the event also allows us to assess the extent telescoping might be distorting the estimates (e.g., Biderman and Lynch, 1981).

- If the respondent reports that it happened more than 4 times, a check is built in to have the respondent confirm that indeed all of these incidents occurred since the anchor date.
- If the respondent does not recall the month in which the incident occurred, she is asked to confirm that the incident indeed took place since the anchor date. If it occurred prior to the anchor date, it is not counted as an incident occurring in the past 12 months.
- If more than one incident is reported in the same month, respondents are prompted to report whether they are part of the same incident or are separate incidents.

References

- Belli, R.F., Shay, W.L. and F.P. Stafford (2001) "Event History Calendars and Question List Surveys: A Direct Comparison of Interviewing Methods" *Public Opinion Quarterly* 65 (1): 45-74
- Biderman, A.D. and J.P. Lynch (1981) "Regency bias on self-reported victimization." *Proceedings of the Section on Social Statistics, American Statistical Association*
- Biderman, A.D., and Cantor, D. (1984). A longitudinal analysis of bounding, respondent conditioning and mobility as sources of panel bias in the National Crime Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 708-713.
- Neter, J., and Waksberg, J. (1964). A study of response errors in expenditures data from household interviews. *Journal of the American Statistical Association*, 59, 18-55.

Question #2. Can race and ethnicity questions be made more consistent throughout the instruments?

Response: We have reviewed the Federal Register Notice dated October 30, 1997 concerning the revision of Statistical Policy Directive No. 15, Race and Ethnic Standards for Federal Statistics and Administrative Reporting. Below we outline our proposal for making adjustments to several items in the instruments to bring them in alignment with the OMB Policy Directive.

Race Questions

We recognize that the question wording varies across the items, as shown in the table below. This variation is necessary to make the question appropriate to the focus of the item (self, other householder, offender) and the mode of the instrument in which it is being asked (mail, CAPI, ACASI, CATI). With these minor variations, we believe that the wording of the questions in each of these items aligns with OMB guidelines. We also believe the response options for the race questions, listed below, are appropriate throughout the instruments.

- White
- Black or African American
- American Indian or Alaska Native
- Asian
- Native Hawaiian or other Pacific Islander

We recognize that the instructions for respondents and interviewers differ by item. We propose to make the instructions for the race questions consistent by using the phrase “Please select one or more.” This is drawn from the following Federal Register Notice statement: “Based on research conducted so far, two recommended forms for the instruction accompanying the multiple response question are ‘Mark one or more ...’ and ‘Select one or more....’ Furthermore, the instruction “select one or more” was approved by OMB in two recent studies conducted by BJS: the RTI mode study and the Westat IVR study.

The proposed edits are displayed in Table 1 below.

Table 1. Proposed edits to race items

Appendix	Instrument (Mode)	Item #	Prior Wording	Revised Wording
C	Mail Roster (Paper)	9	What is his/her race? You may mark more than one.	What is his/her race? Please select one or more.
E	HH Roster (CAPI)	HM011	What is {your/NAME's} race? Please choose all that apply.	What is {your/NAME's} race? Please select one or more.
W3	Demographics (CAPI and CATI)	IQ12	What is your race? Please choose all that apply.	What is your race? Please select one or more.
W	Detailed Incident Form (ACASI)	F4B	What race or races was this person? Was this person... (Mark all that apply)	What race or races was this person? Please select one or more.
W	Detailed Incident Form (ACASI)	F14B	What were the race or races of the persons? Were they... (Mark all that apply)	What were the race or races of the persons? Please select one or more.
V	Detailed Incident Form (CATI)	F4B	What race or races was this person? Was this person...	What race or races was this person? Please select one or more.
V	Detailed Incident Form (CATI)	F14B	What were the race or races of the persons? Were they...	What were the race or races of the persons? Please select one or more.

Ethnicity Questions

We have identified inconsistencies in the way ethnicity is asked throughout the instrument. In some questions, Spanish origin is mentioned, but in others it is not. We will make the wording more consistent, as is shown in Table 2 below, to align with language approved by OMB for the NCVS. Note that some of the questions tailor the wording of Latino or Latina based on the gender of the respondent or the reported gender of the perpetrators. For example, when the respondents are all female, they will be asked if they are Hispanic or Latina. Likewise, if the respondent has reported that the offender was a female, she will be asked if the person who did this to her was Hispanic or Latina.

According to the 1997 OMB directive, the response options for the ethnicity item should be "Hispanic or Latino" and "Not Hispanic or Latino." We will revise all items to provide these response options.

The proposed edits are summarized in Table 2 below.

Table 2. Proposed edits to ethnicity items

Appendix	Instrument (Mode)	Item #	Prior Wording	Revised Wording
C	Mail Roster (Paper)	8	Is this person of Hispanic or Latino origin? Yes/No	Is this person Hispanic or Latino? Hispanic or Latino Not Hispanic or Latino
E	HH Roster (CAPI)	HM0010	{Are you/Is NAME} Hispanic, Latina, or of Spanish origin? Yes/No	{Are you/Is NAME} Hispanic or Latino? Hispanic or Latino Not Hispanic or Latino
W3	Demographics (CAPI/CATI)	IQ10	Are you Hispanic, Latina, or of Spanish origin? Yes/No	Are you Hispanic or Latina? Hispanic or Latina Not Hispanic or Latina
W	Detailed Incident Form (ACASI)	F4a	Was this person Hispanic or Latina/Latino? (wording to be tailored to gender of offender) Yes/No	Was this person Hispanic or Latina/Latino? (wording to be tailored to gender of offender) Hispanic or Latino/Latina Not Hispanic or Latino/Latina
W	Detailed Incident Form (ACASI)	F14a	Were any of the persons Hispanic or Latino? Yes/No	Were any of the persons Hispanic or Latina/Latino? (wording to be tailored to gender of offenders) At least one was Hispanic or Latino/Latina None were Hispanic or Latino/Latina
V	Detailed Incident Form (CATI)	F4a	Was this person Hispanic or Latina/Latino? (wording to be tailored to gender of offender) Yes/No	Was this person Hispanic or Latina/Latino? (wording to be tailored to gender of offender) Hispanic or Latino/Latina Not Hispanic or Latino/Latina

V	Detailed Incident Form (CATI)	F1 4a	Were any of the persons Hispanic or Latino? Yes/No	Were any of the persons Hispanic or Latina/Latino? (wording to be tailored to gender of offenders) At least one was Hispanic or Latino/Latina None were Hispanic or Latino/Latina
---	-------------------------------	----------	---	---

Question #3: What are the plans for publishing the results of the cognitive interview results? Why do you think it is important to publish the results? Is it possible to publish the results without revealing the confidentiality of participants? How does BJS plan on releasing these reports?

There are two reasons why we believe it is important to publish these results. First, it provides the research community a description of the process used to develop the NSHS questions. This type of transparency is important to advance our knowledge of the effects of question design on the measurement of rape and sexual assault. Second, and related to the first reason, these publications will be one of the first qualitative descriptions of how respondents interpret and answer behaviorally specific screening questions. Our experience on the NCVS is that respondent interpretation of screener items is complicated by not only interpretation of particular words, but also the respondent's experiences. The screening items are critical to the scope of the items reported and other researchers can greatly benefit from the qualitative data collected as part of this study. Also note that prior surveys have used the screening items as the primary indicator when classifying as rape and sexual assault. Our results provide information on how a detailed incident form (DIF) supplements that approach when one is interested in counting and classifying events into legally-based categories.

Cognitive interviews are a form of qualitative data collection. Analysis and reporting of these data follow the same practices as qualitative researchers when publishing their results. Results from interviews like this are presented in public meetings and in journals on a regular basis. Scanning almost any program from the annual meeting for the American Association for Public Opinion Research reveals multiple presentations of results of this type.¹ Results will not provide detail that potentially reveals the identity of any of the respondents. For example, idiosyncratic details about respondent situations will not be published. Westat and BJS will review the data included in the papers before they are released.

BJS considers the publication of the cognitive interview results as a fulfillment of its core mission. BJS has invested significant resources in the redesign of the NCVS to improve methodology and increase the survey's value to national and local stakeholders. A section of the BJS website is dedicated to providing information to the public regarding ongoing methodological research in support of the NCVS.

Publication of the NSHS cognitive interview findings will adhere to the standard procedures established and refined by BJS over the last 35 years. As a statistical agency with extensive experience processing and disseminating potentially sensitive information, internal reviews have been developed to insure all statistical research is released in a manner maintaining anonymity and confidentiality as appropriate. Once internal review of the report is complete, we expect to release the findings on the BJS webpage.

¹ For example, in 2011 there was a panel entitled "Insights from Cognitive Interviewing", with papers on results from cognitive interview evaluating the race/ethnicity questions, a web version of the census short form, questions about sexual identity and questions asking about marital status among same-sex couples.

Question #4: Can Westat provide a copy of the full interview from beginning to end including a brief description of what we plan to submit?

Response: Westat has created two new appendices, Z1 and Z2 that provide the entire CATI and ACASI instruments from beginning to end. We have left out the scripts that are used when volunteers and service provider sample are called back to determine eligibility and to schedule their appointments (Appendices M, N and P).

Appendix Z1 has concatenated the following appendices together to demonstrate the flow of the in-person instrument:

- Appendix E: Field ACASI Household Roster
- Appendix Q-2: Field Consent Form ABS and High Risk
- Appendix R: Field Consent Form Service Provider Sample
- Appendix W3: Demographics
- Appendix W1: Event History Calendar
- Appendix W2: ACASI Tutorial
- Appendix W: ACASI Questionnaire
- Appendix U1: Distress and Debriefing

Appendix Z2 has concatenated the following appendices together to demonstrate the flow of the telephone instrument:

- Appendix G: CATI Landline Screener
- Appendix H: CATI Cell Phone Screener
- Appendix S: Phone Consent Landline and Cell
- Appendix T: Phone Consent High Risk Sample
- Appendix U: Phone Service Provider Consent
- Appendix W3: Demographics
- Appendix V: CATI Questionnaire
- Appendix U1: Distress and Debriefing

Question #5: Does the burden estimate include time related to screening the household for selecting the respondent?

Response: Yes. The total burden hours for activities associated with screening the household is estimated to be 3,281 hours. The burden estimates are listed in the Supporting Statement, Section A, page 26, Table 1c. Details are provided in Appendix A – Annual Respondent Burden for the NSHS Interviews – Detail.

The relevant rows from Supporting Statement, Section A, Table 1c are highlighted below.

Table 1c. Respondent activities, times, and respondents for the NSHS Interviews

Task Activity	Sample Type	Time to complete	# of Feasibility Study Respondents	# of Pilot Test Respondents
ACASI Sample				
Read flyer/respond via web or phone	Service Provider High Risk	0.067	72	1,566
Read letters (1 st /2 nd advance, post-survey)	General Population	0.017	394	64,830
Complete hhld roster by mail		0.083	46	7,359
Complete hhld roster in person		0.117	70	13,245

The relevant rows from Appendix A are highlighted below.

Feasibility Test

Table A-1. Assumptions used for calculation of respondent Burden (continued)

Description	Sample Type	Time to Complete Task	Number of responses+	Total expected burden hours*	Total expected burden cost
ACASI Sample (Total)				65.2	\$1,565
Read flyer/respond via web or phone	Service Provider	0.067	24	1.6	\$38
	High Risk		48	3.2	\$77
Read 1 st advance letter and household roster	General Population	0.017	200	3.3	\$80
Read 2nd advance letter and household roster			170	2.8	\$68
Read post-survey letter – eligible households			24	0.4	\$10
Complete household roster by mail	General Population	0.083	46	3.8	\$92
Complete household roster in person	General Population	0.117	70	8.2	\$196

Pilot Study

Table A-1. Assumptions used for calculation of respondent Burden (continued)

Description	Sample Type	Time to Complete Task	Number of responses+	Total expected burden hours*	Total expected burden cost
ACASI Sample (Total)				6,464.5	\$155,020
Read flyer/respond via web or phone	Service Provider	0.067	361	24.1	\$577
	High Risk		1,205	80.3	\$1,926
Read 1 st advance letter and household roster	General Population	0.017	33,072	551.2	\$13,218
Read 2nd advance letter and household roster			28,111	468.5	\$11,235
Read post-survey letter – eligible households			3,827	63.8	\$1,530
Complete household roster by mail	General Population	0.083	7,359	613.3	\$14,706
Complete household roster in person	General Population	0.117	13,245	1,545.3	\$37,055

The activities associated with screening the household and the respective burden are described below.

General population households in the ACASI sample will receive up to three mailings from Westat. The burden for reading each mailing is estimated to be 1 minute. The cumulative burden for these mailings from the Feasibility Study and Pilot Test comes to 1,149 hours ($555 + 471 + 123 = 1,149$).

- a) The initial advance letter and household roster will be mailed to 33,272 households (Feasibility Study = 200 households, Pilot Test = 33,072 households), for a total burden of 33,272 minutes or 555 hours.
- b) The follow-up advance letter and household roster will be mailed to the households that do not respond to the first request, which is estimated to be 28,281 households (Feasibility Study = 170 households, Pilot Test = 28,111 households), for a total burden of 28,281 minutes or 471 hours.
- c) We estimate that 7,405 households will return a completed roster by mail (Feasibility Study = 46 households, Pilot Test = 7,359). We estimate that 52% of these households will have an eligible respondent; these households will receive a letter notifying them that they have been selected for participation. The burden for reading this letter is 3,851 minutes or 64 hours.

The 7,405 households (see above) that return a completed roster by mail, will have an average burden of 5 minutes per household to fill out the roster for a total burden of 37,025 minutes, or 617 hours.

Households that do not respond to the roster by mail will be asked to complete a roster in person. We estimate that 13,315 households will complete the roster in person (Feasibility Study = 70 households, Pilot Test = 13,245 households). The burden to complete the in person roster is estimated to average 7 minutes, for a total burden of 93,205 minutes or 1,553 hours.

Question #6: How are you ensuring that college students are not double sampled – both at home and on college campuses?

Response: In the household screening process, respondents are asked to identify the number of adults ages 18 or older who think of this address as their main home. They are told specifically to exclude “college students who live away from home.” College students will be sampled where they reside at the time of the survey. A separate stratum will be formed of census blocks containing female students in college dormitories or a high proportion of women 18-29 to oversample this high risk group.

Question #7: Why are you proposing a promised incentive for the mail survey when a pre-paid incentive has been shown to be more effective?

We had proposed a promised incentive of \$5 for the mail survey sent to the ABS sample. As suggested by the OMB response, a pre-paid incentive has been found to be more effective than a promised incentive. We had avoided a pre-paid incentive because of prior bad experience with this methodology on another OJP study. In that case, members of Congress had received calls from constituents who objected to the government sending money for a survey. We would like to change to a pre-paid incentive of \$2 because of the proven effectiveness of this methodology. The amount of \$2 has been shown to significantly increase response rates relative to \$1 (e.g., Trussell and Lavrakas, 2004).

References

Trussell, Norman, and P. Lavrakas. 2004. "The influence of incremental increases in token cash incentives on mail survey response: Is there an optimal amount?" *Public Opinion Quarterly* 68, no. 3: 349-367

Question #8: Why is the incentive so high for the main survey given the burden is lower than the surveys cited in the package (NSDUH; MEPS)?

Response: The rationale for the incentive for each of the sample groups is discussed below.

Telephone Interview with RDD sample

The rationale for \$20 is based on studies that have found that promising money over the phone generally requires a significant amount of money to be effective. Our own studies have found \$15 to be a minimum (Cantor, et al., 2006). While there are exceptions to this result, the burden and sensitivity of the questions in the NSHS survey is quite high. The National Intimate Partner and Sexual Violence Survey (NISVS), which asks similar questions to what is on our victimization screener, promises \$10 to initial cooperators, but \$40 to a sample of refusers. The NISVS design acknowledges that it is necessary to use more than \$10 to obtain an acceptable response rate. Our recommended design uses a uniform amount, rather than basing the amount on the difficulty associated with completing an interview. Given the sensitive nature of the survey (see ABS discussion below), the NSHS survey is not conducting any refusal conversion. Consequently, even if differential incentives were desirable, they could not be implemented.

We had proposed a promised (*versus a pre-paid*) incentive for the RDD telephone survey for two reasons. First, it only would be possible to send a pre-paid incentive to about 25% of the sample. Half of the sample will be with persons from the cell-phone RDD frame for which it is not possible to obtain addresses to send an pre-paid incentive. In addition, only 40% to 60%² of the landline phones can be successfully matched to the correct telephone number. This only leaves around 25% of all numbers that can receive a pre-paid incentive.³

A second reason to include a promised incentive is that it allows targeting the person selected to do the interview. A pre-paid incentive can only target a response to increase response to the initial household screening, where anyone in the household can respond. With a promised incentive, the interviewer can target the selected respondent once they are able to talk to her in private. As discussed above we expect that the sensitivity and burden associated with completing the survey is such that a promised incentive will significantly increase the number of women who complete the questionnaire.

ABS in-person interview

OMB has commented that the surveys cited in the original package are significantly longer than the NSHS. To provide a more explicit comparison of the burden, Table 1 provides information for several federally sponsored in-person surveys along dimensions that define survey burden (Bradburn, 1978; Singer, et al, 1999). In the table, burden is defined by the average time to

² This includes those units that are called and it is found that the address does not match the phone number.

³ This was calculated by assuming: 0% of the cell phone sample will have addresses; 50% of landline phones will have an address. With a 50-50 split between the cell-landline sample, this results in $(.5*0\%) + (.5*50\%) = 25\%$ of the sample with usable addresses.

complete the interview, the effort needed to complete the required tasks, the sensitivity of the questions and whether the survey is longitudinal. The footnote in the table provides a key on how the surveys were rated for the 'effort' and 'sensitive' dimensions.

OMB's comment is correct when comparing the size of the incentives and the relative length of each survey. For example, both NSDUH and MEPS are approximately 1 hour, while the NSHS is estimated to average about 19 minutes.⁴ This is true for the other surveys shown in Table 1.

The rationale for an incentive for the NSHS is related to the highly sensitive nature of the survey. We have rated the NSHS as being the most sensitive among these surveys for several reasons. One is the extremely private nature of the topic. This sensitivity leads to a design which does not reveal the specific topic of the survey until the respondent is selected and in a private setting. This is unlike any of the other surveys on sensitive topics such as drug use, use of alcohol or asking about income. Following the practice of other surveys on intimate partner violence, this procedure fosters confidentiality of the topic of the survey within the household. This promotes candid reporting, as well as preventing possible retaliation from other household members. However, revealing the specific topic at this point of the interview introduces additional burden related to the sensitivity of the survey.

Second, the questions have the potential for bringing up negative emotions or feelings. Research on interviews of this type has shown generally that victims of sexual violence find these interviews as a positive experience (Labott et al, 2013; Walker et al, 1997). Nonetheless, they can bring up negative emotions. This aspect of the survey is not unique to other surveys on intimate partner violence, but it is unique among the surveys listed in Table 1.

A third reason the NSHS survey is rated highest on sensitivity is the use of a detailed incident form (DIF). While a relatively small number of respondents will fill out a DIF, this portion of the survey adds burden beyond what similar surveys have done. With one exception (Fisher, 2004), the surveys on intimate partner violence have avoided asking for details because it can be very sensitive. An important goal of the NSHS is to assess the utility of the DIF for purposes of classifying and counting the number of incidents. This is something other surveys have not been able to do cleanly (see response to analysis question). The DIF includes questions on such topics as the type of force that might have been used, the extent alcohol/drugs was involved and how the victim reacted to the situation. This adds significant burden to the task.

A promised incentive plays an important role in motivating respondents to complete the survey. Recent research testing an Interactive Voice Response version of the NCVS found that promising \$10 increased the number of respondents who filled out a victimization screener, as well as completing all of the expected DIFs. In the case of filling out all DIFs, these results found that 30% of respondents did not complete all DIFs without an incentive, while this dropped to 20% for those that received an incentive (Cantor and Williams, 2013). This effect was directly related to the difficulty of the respondent's task. As noted above, we will not be conducting any refusal conversion once the respondent has been informed about the topic of the

⁴ . We note that for those that fill out one DIF, the average will be 30+ minutes. The time to complete will approach 1 hour if 3 DIFs are filled out.

survey. The incentive levels for both ABS and RDD seek to maximize the extent respondents consider participating and completing the survey.

We recognize there is little research that assesses the contribution of burden associated with the qualitative dimension of sensitivity, as compared to burden associated with longer interviews or level of effort (e.g., record keeping). Our proposal of \$40 for the in-person survey assumes the overall burden for NSHS is equivalent to longer surveys such as those listed in Table 1. Nonetheless, the average length of the interview is significantly shorter than the comparable surveys shown in Table 1. We believe an incentive of \$20, equivalent to the proposed telephone version and to the NISVS (see rationale for telephone interview above) is warranted.

High Risk Group

In setting the amount for the volunteer sample we thought the best comparisons were incentives used to recruit participants for cognitive interviews. The success of this component of the study is dependent on recruiting enough individuals to complete 2000 interviews. Our experience with offering significantly lower incentives for cognitive interviews and focus groups is that it will reduce the number of women who volunteer and will have a negative effect on the diversity of individuals who participate on the survey. Young people are typically hard to recruit given all the other competing priorities these individuals have.

As noted in the OMB package, to maintain the random assignment to interview mode, the incentive has to be the same for all high risk volunteers. It is essential that assignment to mode be done once the respondent is deemed eligible and has agreed to be interviewed. This can only be done once the respondent is told what the incentive is.

From our experience with recruiting volunteer samples like this, we believe that it is important to offer at least \$30 to get the attention of a wide diversity of potential respondents.

Service Provider Sample

Like the High Risk group, this group will be asked to volunteer based on flyers handed out at the agencies. For this reason, we recommend an incentive of \$30 for this group as well. This group will also be asked if they want to conduct the interview at the service provider's location or at a place where they can guarantee they can speak confidentially and safely. We are making these special arrangements for this group because of the serious nature of their experiences. If the respondent does travel to do the interview, we propose providing \$10 to offset some of the travel costs they may incur.

Table 1. Incentive and Burden on Selected Federally Sponsored Surveys

Survey	Task	Average Length	Effort	Sensitivity	Panel	Incentive
National Survey on Health and Safety	Auto-biographical questions on sexual assault	19 minutes	Average	High; private information; explicit language; potential of emotional trauma and retaliation	No	\$40
Program for International Assessment of Adult Competencies	Educational Assessments	2 hours	Average	Average	No	\$50
National Epidemiologic Survey on Alcohol and Related Conditions	Auto-biographical questions on alcohol use; provide biological samples	2 hours	Average	Above Average High risk behaviors	Yes	\$90 for interview \$
National Health and Nutrition Survey	Autobiographical Questions on health and physical examination	60 minutes for household interview plus time for exam	High – travel for exam; physical intrusion	Above Average; HIVs; questions on drug use	No	\$90 - \$125 interview, exam Travel reimbursement \$30 - \$50 per phone interview, activity monitor, urine
National Longitudinal Survey of Youth	Autobiographical questions on labor market activities and other life events	65 minutes	Average	Average	Yes	\$40
National Children’s Study	Autobiographical questions on child development	45 minutes	Average	Above Average Personal questions	Yes	\$25
National Health and Aging Trends Study	Autobiographical questions on health and aging	105 minutes	Average	Average	Yes	\$40
Population Assessment of Tobacco and Health Study	Autobiographical questions on tobacco use and health; provide saliva sample	45 minutes	Average	Above Average Risk behaviors	Yes	\$35 for interview \$10 - \$25 per parent interview, bio collection
Medical Expenditure Survey (MEPS)	Autobiographical questions on health expenditures	60 minutes	Above Average;	Above Average	Yes	\$50

			records	Expenditures and income		
National Survey on Drug Use and Health	Autobiographical questions on drug use	60 minutes	Average	Above Average Illegal behavior	No	\$30
ADD Health	Autobiographical questions on health and health related behaviors	90 minutes	Average	Above Average Illegal behavior	Yes	\$40 for latest wave

Effort = Rated as average unless it requires travel, physical procedures or keeping records; Sensitivity – Average unless involves asking about sensitive behaviors (e.g. illegal or high risk) and/or topics that are potentially traumatic experiences; use of explicit language

References

- Bradburn, N. M. (1978). Respondent Burden. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 35–40.
- Cantor, D. and D. Williams (2013) Assessing Interactive Voice Response for the National Crime Victimization Survey. Final Report prepared for the Bureau of Justice Statistics under contract 2008-BJ-CX-K066
- Cantor, David, Holly Schiffrin, Inho Park and Brad Hesse. (2006) “An Experiment Testing a Promised Incentive for a Random Digit Dial Survey.” Presented at the *Annual Meeting of the American Association for Public Opinion Research*, May 18 – 21. Montreal, CA.
- Fisher, B. (2004). *Measuring rape against women: The significance of survey questions*. NCJ 199705, US Department of Justice, Washington DC.
- Labott, S.M., Johnson, T., Fendrich, M. and N.C. Feeny (2013) Emotional Risks to Respondents in Survey Research: Some Empirical Evidence. *Journal of Empirical Research on Human Research Ethics: An International Journal*, Vol. 8, No. 4: 53-66.
- Research Triangle Institute. (2002). 2001 National Household Survey on Drug Abuse: Incentive Experiment. Rockville, Maryland: Project 7190 - 1999-2003 NHSDA.
- Singer, Eleanor, John Van Hoewyk, Nancy Gebler, Trivellore Raghunathan, and Katherine McGonagle. (1999). “The Effect of Incentives on Response Rates in Interviewer-Mediated Surveys”. *Journal of Official Statistics* 15:217-230.
- Walker, E.A., Newman, E., Koss, M. and D. Bernstein (1997) “Does the study of victimization revictimize the victims?” *General Hospital Psychiatry* 19: 403-410.

Question #10: Expand on the analyses that will be done.

Response: The OMB proposal provided a brief outline of the types of analyses the project will complete, but did not go into sufficient detail on how these analyses will contribute to evaluating the methodologies. In this section we describe in more detail the types of analyses that are planned and how they will be used to assess the two instruments. The statistical power is provided for selected analyses.

A primary goal of the analysis will be to evaluate how the approaches implemented on the NSHS improve measures of rape and sexual assault. The analysis will be guided by two basic questions:

1. What are the advantages and disadvantages of a two-stage screening approach with behavior-specific questions?
2. What are the advantages and disadvantages of an in-person ACASI collection when compared to an RDD CATI interview?

The analysis will examine the NSHS assault rates in several different ways. It is expected that comparisons to the NCVS will result in large differences with the NSHS estimates. We also will compare the estimates of the ACASI and telephone interviews. The direction of the difference in these estimates by mode may not be linked specifically to quality. If the rate for one mode is significantly higher than the other, it will not be clear which one is better. We will rely on a number of other quality measures, such as the extent to which the detailed incident form (DIF) improves classification of reports from the screener, the reliability of estimates as measured by the re-interview, the extent of coverage and non-response bias associated with the different modes, and the extent respondents are defining sexual assault differently. In the remainder of this response we review selected analyses to illustrate the approach and statistical power of key analyses.

Comparison of Assault Rates

There will be two sets of comparisons of assault rates. One will be with the equivalent NCVS estimates and the second will be between the two different survey modes.

Comparisons to the NCVS

A basic question is whether the estimates from NSHS, either the in-person or CATI approaches, differ from the current NCVS. The NCVS results are estimates of *incident rates*, that is, the estimated number of incidents divided by the estimated population at risk. However, the sources used in our design assumptions, such as the British Crime Survey, have emphasized lifetime or one-year *prevalence rates*, that is, the estimated number of persons victimized divided by the estimated population at risk. In general, prevalence rates for a given time interval cannot be larger than the corresponding incidence rates. For the NSHS, we expect to estimate both one-year prevalence and one-year incidence rates for rape and sexual assault. Our design assumptions for NSHS focus on one-year prevalence rates and we

anticipate that the NSHS estimates for prevalence will be considerably larger than the incident rates estimated by the NCVS. This assumption is based on prior surveys using behavior-specific questions that have observed rates that differ from the NCVS by factors of between 3 and 10, depending on the survey and the counting rules associated with series crimes on the NCVS (Rand and Rennison, 2005; Black et al., 2011). The use of ACASI and increased controls over privacy for both surveys has also been associated with increasing the reporting of these crimes (e.g., Mirrlees-Black, 1999). Thus, the NSHS design assumes that both the in-person and telephone approaches will yield estimates considerably larger than implied by the NCVS.

Since NCVS estimates are incidence rates, we derived an annual prevalence rate implied by the NCVS we have computed the 6-month rate for women 18-49 years old and doubled it to approximate the 12-month rate. This clearly overestimates the actual rate because it excludes the possibility that some respondents could be victimized in both periods. Table 1 provides these approximate prevalence rates from 2005 to 2011. A multi-year average is used to remove fluctuations due to sampling error related to the small number of incidents reported on the survey for a particular year.

Table 1. Approximate annual prevalence rates and standard errors for the NCVS for females 18-49 during 2005-2011.

Year	Rape, attempted rape, and sexual assault		Rape and attempted rape only	
	NCVS estimate	Standard error	NCVS estimate	Standard error
2005	0.0014	0.0003	0.0011	0.0002
2006	0.0022	0.0004	0.0014	0.0003
2007	0.0025	0.0003	0.0015	0.0003
2008	0.0017	0.0003	0.0011	0.0003
2009	0.0012	0.0003	0.0009	0.0003
2010	0.0020	0.0004	0.0012	0.0003
2011	0.0022	0.0004	0.0016	0.0003
Pooled, 2005-2011	0.0020	0.0001	0.0013	0.0001
Pooled 2005 – 2011 for 5 metro areas*	.0020	.0004	.0013	.0003

Notes: National rates and standard errors derived from the NCVS public use files, downloaded from ICPSR. Pooled rate for 5 metro areas are based on assumption that the rate is the same as the national and the standard error is approximately 3 times as large as the national standard error.

During 2005-2011, the NCVS reflected an annual prevalence rate of approximately 0.0013 for rape and 0.0020 for rape and sexual assault combined (Table 1). Considering both the

NCVS sample sizes in the five metropolitan areas and the effect of the reweighting to reflect the NSHS sample design, the standard errors for the NCVS for the five metropolitan areas combined are likely to be about 3 times as large as the corresponding national standard errors. Thus, the standard error for estimates from the NCVS for rape and attempted rape is likely to be about 0.0003 on the estimate of 0.0013, and the standard error for rape, attempted rape, and sexual assault about 0.0004 on the estimate of 0.0020 (Table 1).

Our design is based on the assumption that the NSHS prevalence rate for rape will be roughly 3 times the NCVS rate and a prevalence rate for rape and sexual assault roughly 15 times the NCVS rate based on the above studies. For design purposes, this translates to assumed prevalence rates for the NSHS of 0.0045 for rape (and attempted rape) and 0.0360 for rape and sexual assault. Table 3 provides the power of comparisons assuming that these differences occur.

Table 2. Expected power for comparisons between expected NSHS Sexual Assault rates with the NCVS

Survey	Rape, Attempted Rape and Sexual Assault			Rape and Attempted Rape		
	Estimate	Standard Error	Power+	Estimate	Standard Error	Power+
NCVS	.0013	.0003		.0020	.0009	
NSHS In-Person	.0045	.0009	90%	.036	.0026	>90%
NSHS: Telephone	.0045	.0011	80%	.036	.0032	>90%

+ Power when compared to the NCVS estimate

For the combined category of rape, attempted rape, and sexual assault, both the in-person and telephone samples easily will yield statistically significant findings when compared to the much lower NCVS results. Estimates of 0.0360 with standard errors of 0.0026 and 0.0032 for in-person and telephone, respectively, are certain to yield significant results when compared to 0.0020 from the NCVS. For the less frequent category of rape and attempted rape, for which an estimate of about 0.0045 is expected, the situation requires a closer check—the standard errors of 0.0009 and 0.0011 for in-person and telephone, respectively, would have associated power of about 90% in the first case and about 80% in the second.

Comparisons between the In-Person and CATI approaches

The OMB package provided estimates of power for comparing the overall rates of rape and sexual assault between the two modes (see Table 3 below). To detect a significant difference in estimates of rape with 80% power, the estimates would have to differ by a factor of 2 (.004 vs. .008). While this large difference is not unusual for many of the comparisons discussed above, it is large when comparing two methodologies that are similar, at least with respect to the questionnaire. Estimates of other sexual assault will be able to detect differences of about 33% of the low estimate (e.g., .03 vs. .04). This drops to around 25% of the estimate when

combined with the high risk sample (Data not shown). So there should be reasonable power for this aggregated analysis.

Table 3. Size of the Actual Difference in Sexual Assault Rates Between Modes to Achieve 80% Power

Type of Assault	Low Estimate in the Comparison	Standard Error of Difference between Modes	Size of Difference to have 80% Power
Rape	.0045	.00146	.044
Other Sexual Assault	.031	.0031	.0095

Use of a Detailed Incident Form

As noted above, the comparison of the rates, while interesting, is not a direct measure of data quality. One of the primary goals of the NSHS is to develop and evaluate a detailed incident form (DIF) to classify and describe events (e.g., see research goal, question #1 above). Prior studies using behavior-specific questions have depended on the victimization screening items to classify an incident into a specific type of event. This methodology relies on the respondent's initial interpretation of the questions to do this classification. For the NCVS, this can be problematic because the screening section does not document the essential elements that define an event as a crime. For example, on the NCVS, a significant percentage of incidents that are reported on the screener do not get classified as a crime because they lack critical elements (e.g., threat for robbery; forced entry for burglary). We are not aware of the rate of 'unfounding' the Census Bureau finds from this process, but in our own experience with administering the NCVS procedures, 30% of the incidents with a DIF are not classified as a crime using NCVS definitions. Furthermore, the screener items may not be definitive of the type of event that occurred. For example, on the NCVS a significant number of events classified as robberies come from the initial questions that ask about property stolen, rather than those that ask about being threatened or attacked (Peytchev, et al., 2013).

The increased specificity of behaviorally-worded screening questions may reduce this misclassification. However, even in this case respondents may erroneously report events at a particular screening item because they believe it is relevant to the goal of the survey, but it may not fit the particular conditions of the questions. Fisher (2004) tested a detailed incident form with behavior-specific questions and found that the detailed incident questions resulted in a significant shift between the screener and the final classification. Similarly, our cognitive interviews found that some respondents were not sure how to respond when they experienced some type of sexual violation, but did not think it qualified for a specific type of question. It may have been an alcohol-related or intimate partner-related event, which the respondent thought was relevant, but did not exactly fit when asked about 'physical force' (the first screener item). Some answered 'yes' to the physical force question, not knowing there were subsequent questions targeted to their situation.

An important analysis for this study will be to assess the utility of a DIF when counting and classifying different types of events involving unwanted sexual activity. This will be done

by examining how reports to the screener compare with their final classification once a DIF is completed. Initially, we will combine both in-person and CATI modes of interviewing for this analysis. This will address the question of whether a DIF, and its added burden, is important for estimating rape and other sexual assault. We will then test whether there are differences between the two different modes of interviewing. We will analyze the proportion of incidents identified in the screener as rape and sexual assault that the DIF reclassifies as not a crime, in other words, unbounding them.

Table 4. Standard errors for the estimated unbounding rates, for true unbounding rates of .10 and .30.

	In-Person		Telephone 18-49		Combined	
	General Populat'n	General Populat'n + high risk	General Populat'n	General Populat'n + high risk	General Populat'n	General Populat'n + high risk
Standard errors of estimated unbounding rate, for unbounding rate=0.10						
Rape	0.06197	0.05277	0.07682	0.06281	0.04823	0.04026
Other Sexual Assault	0.02342	0.01995	0.02904	0.02374	0.01823	0.01522
Total	0.02191	0.01866	0.02716	0.02221	0.01705	0.01423
Standard errors of estimated unbounding rate, for unbounding rate=0.30						
Rape	0.09466	0.08061	0.11735	0.09594	0.07368	0.06150
Other Sexual Assault	0.03578	0.03047	0.04435	0.03626	0.02785	0.02325
Total	0.03347	0.02850	0.04149	0.03392	0.02605	0.02174

Notes: The rates used in the calculations are taken from Table 9 of the October, 2013 submission to OMB. The rates are for purposes of illustration only. The standard errors shown are the standard errors of the estimated unbounding rate given the crime rate.

The results for the illustrative unbounding rates of 0.10 and 0.30 are shown in Table 4. These show that the sample sizes will allow for assessing the overall value of the DIF. For all sexual assaults the confidence intervals for a rate of 10% will be ± 0.03 or less. For example, if the estimate of unbounding is 10%, the study would estimate that the use of the DIF would reduce the rates implied by the screener between 7% and 13%. If the unbounding rate is as high as 30%, then the use of a DIF will reduce the rate from the screener by 25% to 35%.

This should provide the needed perspective on the relative merits of the DIF for classifying events as crimes.

Comparisons of the two modes with respect to relatively low unfounding rates will be able to detect differences of about 10 percentage points with 80% power for all sexual assaults. For example, there is 80% power if one mode has a unfounding rate of 5% and the other a rate of 15%. If the unfounding rates are higher, true differences of around 16 percentage points is required (e.g., 25% vs. 41%) for 80% power.

The confidence intervals for the estimate of rape will be much broader. If the unfounding rate is around 10%, the confidence intervals will be as large as the estimate. Higher rates (e.g. around 30%), which are not likely given the relatively small number reported, will be $\pm 15\%$. For example, if the estimate is 30%, the confidence interval will be between 15% and 45%.

Re-interview and Reliability

A second measure of quality will be estimates of reliability from the 1,000 re-interviews that will be conducted with those who report some type of unwanted sexual activity on the screener. An important element of quality is whether respondents interpret the questions the same way once they are exposed to the entire interview. On the NCVS, for example, there are significant effects associated with repeated interviewing (e.g., Biderman and Cantor, 1984). The extent that respondents change answers is indicative of problems with the instrument. For example, an important question will be whether respondents either do not report an incident at the re-interview or if the incident changes classification. Similarly, change in the dating of the event or which screener items elicited the event provides indications of the stability of estimates.

When examining low rates of inconsistency (e.g., 10%), there will be relatively high power (75%) when the difference between the modes is approximately 10 percentage points (e.g., 10 vs. 20). When computing Kappa statistics, there will be 80% power to detect inconsistencies of approximately 15 percentage points. In other words, there will be sufficient statistical power to detect moderate to large, substantively meaningful, differences in reliability between the modes.

Coverage and Non-Response of Important Subgroups

An important consideration when choosing between an in-person ACASI and an RDD telephone interview is the extent to which there are relative differences in coverage and non-response at the unit level. Differences in non-response are especially important given the continued drop in response rates for RDD surveys. If the differentials are for groups that are thought to be high risk, such as college students or low income groups, there will be a difference in how well the two modes are representing the target population.

The NSHS design will not allow clean separation of the effects of coverage from non-response. However, by comparing the distributions of the final set of respondents, it will be possible to assess their combined effects. If there are differences in the distributions, data from the ACS can be used to assess which survey was closer to the truth.

Table 5 provides the size of the difference needed to achieve 80% power. With 7500 and 4880⁵ interviews in the ACASI and telephone samples respectively, the power for these comparisons will be quite high. For example, estimates as low as 1% in the population will have a standard error for a difference of around .22%. Comparisons of around this percentage would have 80% power when the actual difference was around .7%. As can be seen, proportionately smaller differences would be needed to have 80% power to detect a statistically significant difference.

Table 5. Actual Difference for 80% Power when comparing estimates of Respondent Characteristics

% in Sample	Standard Error of Difference	Actual Difference required for 80% power
1%	0.22%	0.7%
5%	0.48%	1.4%
10%	0.66%	2.0%
20%	0.88%	2.6%
30%	1.01%	3.0%
40%	1.08%	3.2%
50%	1.10%	3.3%

+Assumes a design effect of 1.4 for both the ACASI and Telephone surveys.

Respondent Interpretation of in-scope incidents

The two designs will be compared on how they differ with respect to respondent comprehension of key concepts in the questionnaire. This comparison will be indicative of whether the two designs differ in how respondents may be interpreting the rape and sexual assault questions. At the end of the survey, each respondent will be asked how she would answer the selected screener questions for different scenarios. The scenarios are structured to vary by key dimensions, such as the level of coercion, the type of (non)consent provided and whether alcohol was involved in the situation.

Observed differences between modes could be due to a number of factors, including differential non-response or mode effects (among others). Regardless of the reasons for the differences, this will provide perspective on whether the two designs are leading to noticeably different interpretations of key concepts.

For example, in one scenario, respondents will get one of four different situations related to consent:

⁵ There will be 4880 respondents age 18-49 for the RDD CATI survey.

1. Tom asked if she wanted to have sex. Sue said yes and they proceeded to have sexual intercourse.
2. Tom kissed Sue and they proceeded to have sexual intercourse. Sue did not say anything at the time, but she did not want to have sex.
3. Tom kissed Sue. She tried to push Tom away, but did not actually say no. They proceeded to have sexual intercourse.
4. Tom kissed Sue. Sue said she did not want to have sex, but Tom ignored her and they proceeded to have sexual intercourse. Sue did not resist again because she was afraid Tom would hurt her.

Respondents will be randomly assigned to one of these four conditions. After reading the randomized vignette, respondents are asked a few questions, four after the first vignette and two after the second. Responses to each of the questions can be analyzed separately. The simple analysis will examine the extent the ACASI and telephone responses differ. For example, it will compare the percent of ACASI respondents who report the scenario as a rape when condition 3 is used to those who report it when condition 4 is used.

For purposes of comparing modes, we will restrict the analysis to the ACASI sample and telephone respondents age 18-49, with target sizes of 7,500 and an expected 4,880, respectively. The two modes can also be combined into a sample of over 12,000 for detailed analysis. Because of the randomization, respondents will only receive one from the set of situations, where the number of situations varies from two (involving work) to five (involving drinking behavior and involving relationship). The total sample size is divided among the situations along each dimension. For example, to use 12,000 as the approximate the overall sample size, 3,000 respondents will have answered each of the situations above.

With these sample sizes, the power to detect relatively small differences between modes will be quite high. For example, the third item above, there would be approximately 1,875 ACASI and 1,220 telephone respondents. For example, if the question received a 20% “yes” response, the standard errors of the estimates would be about .009 for ACASI and .011 for telephone. The power to detect differences of 4.4% between the modes would be about 80%. We will also conduct more detailed multivariate analyses that examine differences between modes once accounting for other experimental conditions, such as the respondent’s prior relationship. This should increase the overall precision of the estimates.

References

- Black, M.C., Basile, K.C., Breiding, M.J., Smith, S.G., Walters, M.L., Merrick, M.T., Chen, J., & Stevens, M.R. (2011). The National Intimate Partner and Sexual Violence Survey (NISVS): 2010 Summary Report. Atlanta, GA: National Center for Injury Prevention and Control, Centers for Disease Control and Prevention.
- Biderman, A.D., and Cantor, D. (1984). A longitudinal analysis of bounding, respondent conditioning and mobility as sources of panel bias in the National Crime Survey.

Proceedings of the Section on Survey Research Methods, American Statistical Association, 708-713.

Fisher, B. (2004). *Measuring rape against women: The significance of survey questions*. NCJ 199705, US Department of Justice, Washington DC.

Mirrlees-Black, C. (1999). *Domestic Violence: Findings from a new British Crime Survey self-completion questionnaire*. Home Office Research Study 191, London.

Peytchev, A., Caspar, R., Neely, B. and A. Moore (2013) *NCVS Screening Questions Evaluation: Final Report*. Research Triangle Institute, Research Triangle Park, NC.