O'Muircheartaigh, C., Eckman, S., & Smith, S. (2009). Statistical design and estimation for the national social life, health, and aging project. *Journal of Gerontology: Social Sciences, 64B(S1)*, i12–i19, doi:10.1093/geronb/gbp045.

Statistical Design and Estimation for the National Social Life, Health, and Aging Project

Colm O'Muircheartaigh,[1,2] Stephanie Eckman,[2] and Stephen Smith[2]

[1]Harris School of Public Policy and [2]National Opinion Research Center, University of Chicago, Illinois.

Objectives. The paper discusses the sample design of the National Social Life, Health, and Aging Project (NSHAP) and how the design affects how estimates should be calculated from the survey data. The NSHAP study allows researchers to study the links between sexuality and health in older adults. The goal of the design was to represent adults aged 57–85 years in six demographic domains.

Methods. The sample design begins with a national area probability sample of households, carried out jointly with the 2004 round of the Health and Retirement Study. Selection of respondents for NSHAP balanced age and gender subgroups and oversampled African Americans and Latinos. Data collection was carried out from July 2005 to March 2006.

Results. The survey obtained an overall response rate of 75.5%.

Discussion. The complex sample design requires that the selection probabilities and the field implementation be accounted for in estimating population parameters. The data set contains weights to compensate for differential probabilities of selection and response rates among demographic groups. Analysts should use weights in constructing estimates from the survey and account for the complex sample design in estimating standard errors for survey estimates.

Key Words: Design effect—Health—Sample design—Sample size—Sexuality.

THE overall goal of the National Social Life, Health, and Aging Project (NSHAP) is to study the links between health and sexuality in the lives of older Americans. The sample and estimation plan that we developed was intended to generate an observed sample that can be generalized to the target population: U.S. adults aged 57–85 years living in households. The foundation of the design is a probability sample that gives each element in the population a known nonzero chance of being selected into the sample. The NSHAP sample is based on a standard multistage area probability design, which selects large area units at the first stage, smaller area units at the second stage, and households at the third stage. National Opinion Research Center (NORC) entered into a partnership with the Survey Research Center at the University of Michigan, and together they conduct the Health and Retirement Study (HRS) for these stages of the design, an arrangement which benefited both surveys. Selection of eligible members from these households used an innovative sampling technique to meet a variety of design goals and constraints. The next two sections of this paper give details on the NSHAP sample design.

After selection of individuals for the NSHAP survey, NORC interviewers visited each case to complete a face-to-face interview and conduct biomeasures using a modularized questionnaire design to reduce respondent burden. We achieved a response rate of 75.5%. Below we discuss the fieldwork and its effect on estimation in more detail.

We discuss the details of the NSHAP sample design not only because it provides important ideas for future surveys of similar populations but also because an understanding of the design is crucial in calculating appropriate estimates and their standard errors. The final two sections of the paper discuss the implications of the design for estimation procedures.

Sample Design

The NSHAP sample consisted of multiple stages of selection: (a) two area stages, in which geographic areas were selected into the sample with probabilities proportional to their sizes, (b) a household selection stage in which a sample of households was selected from the selected areas for screening, (c) an individual selection stage in which persons were selected for the NSHAP interview. These stages together determine the probabilities of selection of the individuals in the study. This design is a classic multistage area probability sample (for details on this class of sample designs, see Harter, Eckman, English, & O'Muircheartaigh, in press).

NSHAP wished to interview adults aged 55–85 years. However, only approximately 30% of U.S. households contain an individual in this age range. Identifying such households and the eligible individuals within them would have involved an extremely expensive (about $2 million) and time-consuming screening of a large sample of households. At the time that we were planning the NSHAP design, the HRS (also funded by the National Institute on Aging) was about to embark on the recruitment of a new cohort. Through an innovative collaboration between NSHAP and HRS (and between the NORC and the Institute for Social Research [ISR], the respective survey organizations), the screening for both surveys was carried out as a single operation, with

substantial saving in costs. As HRS interviewers screened households in the selected segments for individuals eligible for their survey, they also identified individuals who were eligible for NSHAP. HRS screening took place from February to November 2004. At the end of their data collection period, they sent all NSHAP-eligible individuals to NORC, and we selected our final sample of households and individuals from this database. This sharing of field resources allowed NSHAP to have a much larger sample size than would otherwise have been possible. However, this collaboration did require that the NSHAP redefine its target population to adults aged 57–85 years. This change meant that the HRS and NSHAP populations were nearly nonoverlapping.

Coverage of the Sampling Frame

Undercoverage occurs whenever some eligible persons have no chance to be selected for the survey. There are two minor sources of undercoverage in the NSHAP design: First, as the survey was to be carried out as a household survey, the population was limited to adults living in households; thus, the institutionalized population and the homeless were excluded; second, those absent from the country during the period of the fieldwork were excluded.

The only substantial source of undercoverage arises from the link between HRS and NSHAP fieldwork. HRS was recruiting the 50- to 56-year-old cohort (the Early Baby Boomers) and their partners as well as the next cohort, the 44- to 49-year-old Middle Baby Boomers and their partners. Due to HRS's complex eligibility rules, 57- to 85-year-olds living in households with 44- to 56-year-old *nonpartners* would not be available for NSHAP by virtue of their residence in the household of an HRS-eligible individual.

To estimate the magnitude of the undercoverage due to the loss of these nonpartners, we used the household composition data in the U.S. Census Bureau's Public Use Microdata Set. Overall, we estimated that this constraint could exclude just more than 5% of the NSHAP-eligible population (6% of the eligible women and 4% of the men). This is a relatively low degree of undercoverage overall, but there is some concern that those excluded would be concentrated in particular (and potentially interesting) subclasses of the population. The most common reason that we expected individuals to be excluded was that they were living with an adult child. Other common reasons included living with a sibling, an unmarried partner, or an unrelated housemate. Data about the excluded cases from the HRS recruitment process itself are not available.

Area Stages of the Design

The sample designs for NSHAP and HRS are identical at the area stages. The first stage consists of primary sampling units (PSUs; either metropolitan areas or counties) selected with probability proportional to size. Within selected PSUs, second stage units (segments) were formed from Census blocks and selected with probability proportional to size. In order to generate sufficient sample size for African American and Latino subsamples, the probabilities of selection of segments with more than 10% African American or Latino population were more than doubled relative to all other segments. This meant that all adults living in these segments, whether African American or Latino, or not, were overrepresented in the sample at this stage. The spatial correspondence between the HRS and NSHAP samples also has significant potential for future joint analyses of the data from the two surveys. For details on selection of multistage area probability samples generally, see Harter et al. (in press). At the time of this writing, some details on the HRS 2004 design were available at the HRS Web site at the Institute for Social Research, University of Michigan; see http://hrsonline.isr.umich.edu/.

Within these selected segments, a full listing of housing units (households) was carried out by HRS field staff. Health and Retirement Study interviewers selected 30,000 households from those listed. Although they planned to select households with equal probability within four domains defined by the concentration of minorities, they deviated from this plan and oversampled segments, regardless of domain, that were found to have higher proportions of eligible persons. HRS also subsampled cases to hasten the end of the fieldwork (private correspondence from HRS statisticians). ISR then delivered all screened households containing at least one NSHAP-eligible member (except as discussed earlier), and we performed additional stages of selection.

Design for Sample of Households and Individuals Within Households

In planning the analyses we wished to run with the NSHAP data, we indentified six domains of interest, that is, subclasses of the population for which separate estimates would be required: three age groups, each subdivided by gender. We determined, using approximate power calculations, that a sample size of 500 would be required for each subclass, giving an overall sample size of 3,000. The binding constraints are those for men and women in the oldest age group. Although we did not use race or ethnicity in the formation of our explicit domains, another design goal was to overrepresent African Americans and Latinos in our final sample.

A general principle of estimation is that, ceteris paribus, a sample design in which individuals are selected with equal probabilities will provide more precise estimates (estimates with smaller standard errors) than a sample in which the probabilities vary arbitrarily from individual to individual (Kish, 1965, 1992). Thus, in selecting households and individuals into the study from the frame provided to us by the HRS screening, we wished to equalize the probabilities of selection of the individuals as much as possible within our six domains.

To avoid possible within-household contamination of responses, we also wanted to select no more than one person in any household. This decision made it much more difficult to meet our domain targets as the number of respondents available for selection into a particular sample was thereby much reduced. For example, a household containing an 82-year-old woman and an 84-year-old man contains an individual from our two challenging domains, but it was not possible to have both of them in the sample.

The screener data delivered by ISR contained 7,407 households, each of which contained at least one person born before 1948, together with data on adults within these households. Sample eligibility for NSHAP was defined based on year of birth (1920–1947 inclusive). In all, 6,974 of the delivered households had at least one person born in this range (9,816 eligible people). Twenty-three percent of the eligible people were identified as African American and/or Latino. This list (6,974 households/9,816 people) constituted the sampling frame for the selection of individuals for NSHAP.

Frame preparation.—The HRS interviewers attempted to collect name, race/ethnicity (race was collected in the HRS screener instrument as a dichotomous variable: minority [meaning African American or Latino] and nonminority [all others]), and birth year for all eligible individuals in the households they screened. The data quality was imperfect, with design and estimation implications. Several steps were necessary to prepare the frame for the sample selection process: gender coding, gender and race imputation, and subsampling in segments oversampled by HRS due to race/ethnic composition.

The HRS screening operation did not collect gender. Because gender was so important to the NSHAP sample design, we coded gender for each age-eligible case from name and family relationship data. In conducting the household roster to identify individuals eligible for NSHAP and HRS, interviewers permitted respondents to identify household members by initials or family roles ("husband," "abuela") rather than by names. Tourangeau, Shapiro, Kearney, and Ernst (1997) find that this can reduce undercoverage in rosters. We coded 87.52% of the eligible cases (12.48% contained no data from which we could deduce gender) and 52.09% of these were determined to be women.

We imputed gender for the 12.48% of cases where it could not be determined, and we imputed race/ethnicity for the 1.23% of cases where it was missing. Age was not missing for any cases on the frame. For gender, imputation was done systematically, sorted on PSU, segment, and individual within segment. After this step was completed, the eligible individuals consisted of 52.10% women. Imputation of race/ethnicity was based on the dominant race of the segment. In 19 of the 416 segments, African American/Latino individuals were the dominant group and all cases with missing race/ethnicity were imputed to this category. In all other segments, cases with missing race/ethnicity data were imputed to "not African American/Latino." After this step was completed, 22.61% of eligible individuals were coded as African American/Latino. Gender and race imputation was carried out only to form strata for use in the sample selection process: The final sample file does not include these imputed variables.

The HRS national sample design oversampled segments with high minority concentration and household within these segments, introducing unequal probabilities of selection for all residents in these segments. Our design intention, however, was to increase the selection probabilities only for African American and Latino individuals. To reduce diversity of selection probabilities of nonminorities in these segments (and thus produce a sample that has more nearly equal probability), we subsampled these cases. Prior to the selection of households for interview, we selected a sample of nonminority individuals (not households) and discarded them from the frame. (Because additional subsampling was carried out by ISR during the screening, this adjustment did not fully equalize the selection probabilities among the nonminority cases.)

After imputation and subsampling, the final frame consisted of 7,768 eligible individuals in 5,920 households (average of 1.31 eligible members per household). In total 51.66% of the eligible individuals were women and 28.57% were racial/ethnic minorities. Individuals were coded into three age categories based on year of birth: 49.11% of eligible individuals were in the first age category (57–65 years; 1939–1947), 29.70% were in the second age category (66–74 years; 1930–1938), and 21.19% were in the last age category (75–84 years; 1920–1929).

Size of sample.—Our objective was to complete interviews with 3,000 eligible respondents, with approximately 500 completed interviews in each of the six age/gender domains. We anticipated a 5% ineligibility rate (although the sample had been recently screened, we did expect some loss due to moving or death) and a response rate of 70% or a little more. Thus, the necessary number of cases to issue to the field staff was $3{,}000/.95/.7 = 4{,}500$. To optimize representation, we felt that we should maximize the response rate; consequently, we decided to select 4,400 individuals from the frame; to generate 3,000 interviews under our assumption of 5% ineligibility, this would require a response rate of 71.7%.

Sample Selection

The objective was to draw a sample of 4,400 people with equal sample sizes in the six target domains. The constraints we faced were to select only one individual per household and as much as possible equalize selection probabilities within domains. The overall selection probability

Table 1. Final Sample Counts (race and gender imputed where necessary)

	Female				Male				
	Age 57–65	Age 66–74	Age 75–84	Female total	Age 57–65	Age 66–74	Age 75–84	Male total	Grand total
African American/Latino	252	188	167	607	253	225	140	618	1,225
Not African American/Latino	547	506	552	1,605	610	513	447	1,570	3,175
Grand total	799	694	719	2,212	863	738	587	2,188	4,400

of a case is the product of the probabilities at each stage: PSU, segment, household (including subsampling by HRS) subsampling of nonminority cases after screening, and selection of households and individuals within households for interview. Our goal in selecting households and individuals for NSHAP was to manipulate these last two probabilities such that the overall probability of selection was nearly constant across all cases in each domain; the closer a sample is to having equal probabilities (referred to as an equal probability of selection method sample), the higher the precision, ceteris paribus. We devised an iterative process that would select individuals with probabilities as close to these ideal probabilities as possible; details are given in Appendix A.

Result

The final sample consisted of 4,400 individuals, distributed as shown in Tables 1 and 2. We were not able to achieve equal numbers of selections in each of the six age–gender cells, but we came as close as we could given the fixed size of the screened sample and our additional constraints. In total, 12.55% of the final sample had missing gender; 0.89% had unknown race. In total, 18.35% had first and/or last name missing (13.45% had first name missing). (In Tables 1 and 2, cases with unknown gender or race are shown by their imputed values.)

Data Collection Issues

At each stage of implementation, there are deviations from the optimal or intended execution. These arise from nonresponse in the screener, nonresponse in the interview, and other random deviations from the expected outcomes. Whenever feasible, we provide weights to compensate (in part at least) for these deficiencies.

Screener Response

The HRS screened approximately 30,000 households. The overall screener completion rate was 95%. As the NSHAP sample used the screened sample as a frame, this 5% shortfall is essentially noncoverage for NSHAP. We do not know the NSHAP eligibility rate among these unscreened households. We do not make any adjustments to the weights for this undercoverage, which is equivalent to assuming that the unscreened cases are identical to the screened cases.

Modularization of the Questionnaire and Biomeasures

As discussed in Smith and colleagues (2009), we conducted a pretest of our questionnaire and interviewing methods. The pretest showed that collecting all the data on each respondent would make the interview unacceptably long and would very likely seriously compromise the NSHAP response rate. To reduce respondent burden to an acceptable level while obtaining population-representative data on as many key variables as possible, a modular implementation was designed. All respondents received a set of core interview and biomeasure items; the remaining measures were allocated to two questionnaire modules and three biomeasure modules. Respondents were randomly assigned to one of six paths containing a set of these modules. The paths and their contents are described in Smith and colleagues. Respondents assigned to paths where modularized interview questions were not asked were instead given a mail-in self-administered questionnaire that included these items. As a check on the fidelity of implementation of the randomization, Table 3 shows the number of completed interviews for each of the six paths. A χ^2 test of significance yields a value of $\chi^2 = 3.707$ with 5 *df*; $p = .5$ (not significant).

The modularization affects the design (and consequently the estimation) in two ways. First, the number of cases for which data are available varies depending on whether a measure was included in the core or only in one or more

Table 2. Final Sample Percentages (race and gender imputed where necessary)

	Female				Male				
	Age 57–65	Age 66–74	Age 75–84	Female total	Age 57–65	Age 66–74	Age 75–84	Male total	Grand total
African American/Latino	5.73	4.27	3.80	13.80	5.75	5.11	3.18	14.05	27.84
Not African American/Latino Black/Hispanic	12.43	11.50	12.55	36.48	13.86	11.66	10.16	35.68	72.16
Grand total	18.16	15.77	16.34	50.27	19.61	16.77	13.34	49.73	100.00

Table 3. Random Assignment of Modules

Path (all paths contain the core questionnaire, core biomarkers, and core leave-behind questionnaire)	No. of cases	Test
1 Modules A, B, and C	527	$\chi^2 = 3.707$ with
2 Modules A, B, and E	511	5 *df*; $p = .5$
3 Modules A, B, and E; module B as leave behind	475	(not significant)
4 Modules A, C, and E; module B as leave behind	511	
3 Modules B, C, and D; module A as leave behind	497	
3 Modules B, C, and E; module A as leave behind	484	
All paths	3,005	

Table 4. Overall and Domain Response Rates

	Weighted response rate (%)
Overall	75.5
Age 57–65	78.6
Age 66–74	73.9
Age 75–86	70.7
Nonurban	77.5
Urban	73.7
Women	74.8
Men	76.3
Minority (HH level)	79.5
Nonminority (HH level)	75.1
Minority Status Not Known by HH informant (HH level)	34.9
Minority refused (HH level)	44.1

Note: HH = household.

of the modules; the expected variance is essentially inversely proportional to the number of cases, and thus, some estimates will have larger variances than others simply because they were asked of fewer cases. Second, for a number of questionnaire items, some respondents will have received the items in a face-to-face interview, whereas others will have completed them in a self-completed questionnaire. Whenever data collection modes are mixed in this way, the possibility of a mode effect arises: The mode of collection may influence the data reported by the respondent. (A review of the extensive literature on mode effects is outside the scope of this paper, but see AAPOR, 2008.)

Response Rates and Nonresponse Bias

A survey response rate is defined as the rate of successful completion of interviews. It is generally interpreted as a measure of how successfully the responding cases represent the population. In the case of an equal probability sample design, this is simply the percentage of selected eligible cases for which an interview is obtained. In the case of a sample where selection probabilities vary across different domains in the population, the definition of the response rate is more complex, and the simple unweighted rate is inappropriate. In estimating a characteristic of the population, survey data should be weighted to take into account probabilities of selection; otherwise, the estimate would overrepresent cases with low probabilities of selection. The calculation of response rates is no different. The American Association for Public Opinion Research (AAPOR) provides a document that explains in detail how response rates should be calculated for different modes and sample designs (Groves, 2006; Groves & Couper, 1998).

Table 4 presents the weighted response rates, using AAPOR's RR2, for the survey as a whole and for selected domains. Although we do not show the unweighted response rates, they are very similar to the weighted rates, reflecting the relative uniformity of the response rates across the different domains of the sample design.

Whenever a survey fails to interview all eligible cases, there is potential for nonresponse bias. Bias can be introduced into survey estimates when the nonresponding cases are different from the responding cases. Understanding, preventing, and adjusting for nonresponse is an active area of research in the survey methodology field, and a discussion of this literature is not possible here; Kalton and Kasprzyk (1986) provide a review.

We can examine response rates across domains to gauge the risk of nonresponse bias in the measures collected by NSHAP. We see in Table 4 that response rates vary 8% points across the three age groups, four points between the urban and nonurban groups, and only 2.5 points between men and women. There is a 4.5-point difference in response rates for minority and nonminority households (and much larger differences between these and households of unknown minority status, though there are few cases in these cells). We find these results somewhat reassuring: they suggest that demographically the respondents and the nonrespondents seem to be similar. In a later section, we discuss how we adjusted the weights to account for nonresponse.

The earlier discussion has focused on unit response rates, which measure response to the survey as a whole, but we are also concerned with response rates and nonresponse bias on specific items or sets of items. For the face-to-face interview, item response rates (again weighted to account for differential selection probabilities) were uniformly high. The response rate to the mail-back survey was satisfactory at 83%.

The biomeasures, similarly, had commendably high response rates, but we believe that this set of items might show more nonresponse bias than others; those who agreed to the biomeasures may have different health status than those who refused them outright. We suggest that analysts consider imputing biomeasure results for those cases that refused them before generating population estimates based on the observed cases (Jaszczak, Lundeen, & Smith, 2009). For more details on the collection of the biomarker variables in the NSHAP interview, see Kalton and Kasprzyk (1986).

Table 5A. A Priori Classification of Respondents (based on screener data)

Table of age category by gender (based on preselection information)			
	Gender		
Age category	Female	Male	Total
Age 57–65			
Frequency	581	616	1,197
Row Pct	48.5	51.5	
Col Pct	38.6	41.1	
Age 66–74			
Frequency	470	515	985
			32.78
Row Pct	47.7	52.3	
Col Pct	31.2	34.4	
Age 75–86			
Frequency	456	367	823
			27.39
Row Pct	55.4	44.6	
Col Pct	30.3	24.5	
Total	1,507	1,498	3,005
	50.2	49.8	100

Note: Row Pct = Row Percent; Col Pct = Column Percent.

Table 5B. A Posteriori Classification of Respondents (corrected with interview data)

Table of age category by gender (corrected after selection)			
	Gender		
Age category	Female	Male	Total
Age 57–65			
Frequency	597	609	1,206
Row Pct	49.5	50.5	
Col Pct	38.5	41.9	40.1
Age 66–74			
Frequency	476	487	963
Row Pct	49.4	50.57	
Col Pct	30.7	33.47	32
Age 75–86			
Frequency	477	359	836
Row Pct	57.1	42.94	
Col Pct	30.8	24.67	27.8
Total	1,550	1,455	3,005
	51.6	48.4	100

Note: Row Pct = Row Percent; Col Pct = Column Percent.

Sample Outcomes

In all, we completed interviews with 3,005 selected respondents, just above our target of 3,000. Tables 5A and 5B give the numbers of interviews achieved in each of the six age/gender domains. Table 5A presents the a priori classification of respondents based on the screener data (imputed where necessary, see above); Table 5B presents the a posteriori classification (corrected with interview data where possible). The two tables are quite close, though we see a slight tendency to misclassify more women than men and to understate the numbers in the highest and lowest age classes.

Weighting

The complex design of NSHAP requires that the data be weighted in the analysis in order to provide unbiased estimates of population characteristics. We provide two weight variables to meet the needs of researchers. The steps in the construction of the weights are given subsequently.

Baseweight

The selection probability for each case in the NSHAP sample is the product of its household's probability of selection for the HRS screening operation (π_{HRS}) and its probability of selection for the NSHAP survey, given its selection for HRS (π_{NSHAP}). The first probability includes the probability of selection of the PSU and segment that contain the case and the probability of selection of the household within the segment, as well as any subsampling adjustments late in the HRS field period. (As mentioned earlier, HRS carried out some modest subsampling of cases both to improve the hit rate of HRS-eligible households and to speed up the close of data collection.) The components of the second probability are any adjustments due to the subsampling of nonminority cases in minority segments, the probability of selection of the household for NSHAP, and the probability of selection of the given individual within the household. The baseweight is the inverse of the overall unconditional probability of selection: baseweight = $(\pi_{HRS} \times \pi_{NSHAP})^{-1}$.

Adjustment for eligibility at the time of interviewing.— Not all screened cases were truly eligible for the NSHAP survey; some were outside the eligible age range, others had moved out of the study population; the ineligible cases are not included in the final data set. The weights for eligible cases were unchanged in this step. (No cases were finalized with unknown eligibility status.)

Adjustment for Nonresponse

Nonresponse of any magnitude threatens the basis of inference from the survey data to the population. We provide an adjustment to the weights to account for nonresponse. All nonresponse adjustments rely on a model that makes assumptions about the nonrespondents. The method we used, which Kalton and Kasprzyk (1986) call sample-based weighting, assumes that once we control for a few key characteristics, nonrespondents are like respondents. The only variables that can be used to control for nonresponse are those that exist on both the responding and the nonresponding cases. Age and race/ethnicity provided the greatest discrimination in response rates (see Table 4). In each of the six cells formed by crossing age and race/ethnicity, weights for responding cases were increased by the reciprocal of the cell-level response rate such that the responding cases take on the weight of the nonresponding cases. To the extent that the correspondence between respondents and nonrespondents is closer within these adjustment classes than it is overall, adjusting the weights separately within these classes will improve the validity of our estimates (Kish, 1992).

Table 6. Some Examples of the *deff* for NSHAP Data

Variable	*deff*[a]
Do you currently have a romantic, intimate, or sexual partner?	1.02
How is your sense of smell? 5-point scale	1.50
How many living grandchildren do you have?	2.37
How many living sons do you have?	2.38
Did you attend college or university?	3.88

Notes: *deff* = design affect; NSHAP = National Social Life, Health, and Aging Project.

[a]This effect incorporates all stages of selection, for both the Health and Retirement Study screener sample and the NSHAP design.

Scale Adjustment

There are two sets of weights provided with the final NSHAP data set: WEIGHT and WEIGHTNR; WEIGHTNR includes the nonresponse adjustment and WEIGHT does not. Both weight variables were rescaled so that they sum to the total number of completed interviews.

Use of Weights and Calculation of Standard Errors

Use of Weights

We recommend that all analyses carried out with NSHAP data incorporate weights. At a minimum, the weight variable without the nonresponse adjustment (called WEIGHT in the data set) should be used; otherwise, the estimates will not represent the population and may be subject to serious biases (Kish & Frankel, 1974). We suggest that in general, the weights incorporating the adjustment for nonresponse (WEIGHTNR) are to be preferred over the weights without the adjustment. These adjusted weights help ensure that estimates project to the known structure of the selected sample.

Calculation of Standard Errors

Although using weights will ensure that analysts have the right point estimates, the standard errors (and confidence intervals [CIs]) on these estimates will be incorrect unless additional care is taken. To calculate standard errors correctly for NSHAP data, it is necessary to take into account the sample design and the fieldwork outcomes. Importantly, failing to account for the design will lead to serious *underestimation* of standard errors and CIs (Kish, 1965; Lee & Forthofer, 2006; Verma, Scott, & O'Muircheartaigh, 1980). Note that ignoring the design (and underestimating CIs) is the default behavior in most statistical packages, which will lead researchers to conclude that their results are statistically significant when they are not.

Three aspects of the sample design can have a substantial effect on standard errors: *stratification*, *clustering*, and *unequal probabilities of selection*. The design effect (*deff*) summarizes the combined effect of these three influences on the variance of estimates from a sample. The square root of the *deff*, called the design factor (*deft*), gives the effect of the design on standard errors. A *deff* on a given estimate less than one indicates that the estimate from a complex sample is more efficient (has lower variance and standard error) than one from a simple random sample of the same size. A *deff* greater than one indicates that a complex sample gives less efficient estimates. Stratification tends to reduce the *deff* and clustering, and unequal weights tend to increase it. Different variables within a given survey will have different *deff* values because some are more highly clustered than other: Variables with high rates of within-cluster homogeneity suffer more (have a higher *deff*) than those that have low rates of homogeneity (Kreuter & Valliant, 2007).

Table 6 presents estimates of *deffs* for a number of variables. The estimates in Table 6 were calculated using the nonresponse-adjusted weight, WEIGHTNR. Here we can see that whether respondents have a sexual partner is not homogenous within clusters (whether you have a sexual partner is not related to whether your neighbor does, for the NSHAP population). The *deff* for this variable is very close to 1.0, and our sample design is just about as efficient as a simple random sample of the same size in estimating this variable. Conversely, there is a high degree of within-cluster homogeneity in education: People who live together in clusters tend to have the same levels of education. This clustering in the variable means that our design is much less efficient at estimating this characteristic than an unclustered sample would be. CIs on estimates of the proportion of the NSHAP population that has a college education will be nearly twice as large ($\sqrt{3.88} = 1.97$) as those from a simple random sample of the same size. This discussion has implicitly assumed that the only quantities being estimated are population means.

Correctly estimating standard errors requires passing stratum, cluster, and weight variables into appropriate statistical software. Most packages offer special routines for this kind of estimation: Stata's svy commands, R's survey package, and SAS's proc surveymeans, proc surveyfreq, etc. The NSHAP data file includes stratum and cluster identifiers as well as weights so that these can be passed into the software routine to produce appropriate standard errors.

Conclusions

The paper describes the design and implementation of the sample for the NSHAP. The sample design began with a national area probability sample of households carried out jointly with the HRS. Subsequently, the selection of respondents for NSHAP produced a balanced sample across age and gender subgroups, with an oversample of African Americans and Latinos. The sample equalized as far as possible the probabilities of selection of individual respondents, given the overall constraints. The complex nature of this design requires that analysts use weights that produce unbiased estimates of the population parameters. The data file contains two weight variables that enable analysts to compensate for

the differential probabilities of selection of individuals and the differential response rates for identifiable subclasses of respondents. The analysis of the data should further take into account the stratified and clustered nature of the design to produce unbiased estimates of standard errors for the survey estimates; these variables are also available on the NSHAP data set.

The sample we designed and implemented for NSHAP succeeded in achieving the goals of NSHAP subject to the constraints imposed by the sometimes conflicting objectives. Partnering with HRS enabled us to obtain a larger sample for the same data collection budget than we would have been able to achieve without this partnership. Such partnerships should be considered by other major studies that require substantial screening efforts to identify special target populations.

Funding

The NSHAP is supported by the National Institutes of Health—the National Institute on Aging, Office of Women's Health Research, Office of AIDS Research, and the Office of Behavioral and Social Science Research (5R01AG021487).

Acknowledgments

We also wish to thank E. Scheib for her assistance in carrying out the NSHAP sample design. All authors were actively involved in the statistical design of the NSHAP and all contributed to the conceptualization of the manuscript. C.O'M. conceived the statistical design of NSHAP and wrote the manuscript with the assistance of S.E. S.E., C.O'M., and S.S. were involved in all stages of its implementation, which was directed by S.S. The response rate calculation and weighting was carried out by S.E. All authors participated in editing the manuscript for intellectual content.

Correspondence

Address correspondence to Colm O'Muircheartaigh, Harris School of Public Policy, University of Chicago, 1155 East 60th Street, Chicago, IL 60637. Email: colm@uchicago.edu

References

American Association for Public Opinion Research. (2008). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (5th ed.). Lenexa, KS: American Association for Public Opinion Research.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, *70*, 646–675.

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

Harter, R., Eckman, S., English, N., & O'Muircheartaigh, C. (in press). Applied sampling for large-scale multi-stage area probability designs. In P. Marsden & J. Wright (Eds.), *Handbook of survey research* (2nd ed.). New York: Elsevier.

Jaszczak, A., Lundeen, L., & Smith, S. (2009). Using non-medically trained interviewers to collect biomeasures in a national in-home survey. *Field Methods*, 21(1), 26–48.

Kalton, G., & Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, *12*, 1–16.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Kish, L. (1992). Weighting of unequal pi. *Journal of Official Statistics*, *8*, 183–200.

Kish, L., & Frankel, M. R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society, Series B*, *36*, 1–22.

Kreuter, F., & Valliant, R. (2007). A survey on survey statistics: What is done and can be done in stata. *Stata Journal*, *7*, 1–21.

Lee, E. S., & Forthofer, R. N. (2006). *Analyzing complex survey data* (2nd ed.). Thousand Oaks, CA: Sage.

Smith, S., Jaszczak, A., Graber, J., Lundeen, K., Leitsch, S., Wargo, E., & O'Muircheartaigh, C. (2009). Instrument development, study design implementation, and survey conduct for the national social life, health, and aging project. *Journals of Gerontology: Social Sciences*, 10.1093/geronb/gbn013.

Tourangeau, R., Shapiro, G., Kearney, A., & Ernst, L. (1997). Who lives here: Survey undercoverage and household roster questions. *Journal of Official Statistics*, *13*, 1–18.

Verma, V. K., Scott, C., & O'Muircheartaigh, C. (1980). Sample designs and sampling errors for the world fertility survey (with discussion). *Journal of the Royal Statistical Society, Series A*, *143*, 431–473.

Appendix A: Details on Selection of Households and Individuals for NSHAP

To select households and individuals for the NSHAP study, we aimed not only to equalize probability of selection with our six age and gender domains but also to select only one person per household. We devised an innovative iterative process to come as close as we could to equal probabilities while not exceeding our constraint.

We first calculated the ideal probability of selection for each person in the frame: that which would lead to equal probability samples within each domain. But these probabilities could not be attained while satisfying our constraint of selecting just one person per household. We treated these ideal probabilities as measures of size (mos) for each individual in the frame.

Let mos_i be the measure of size for person i; the mos of household h is $\text{mos}_h = \sum_{i \in h} \text{mos}_i$. We then rescaled these mos_h so that the sum across all households equaled the total number of households (5,920): $\text{mos}'_h = \frac{\sum \text{mos}_h}{5{,}920} * \text{mos}_h$.

At this point, some households had measures of size mos'_h that were too large. Because we wanted to select 4,400 households of 5,920 (a sampling interval of 5,920/4,400 = 1.345), households where $\text{mos}'_h > 1.345$ would have a chance to be selected twice. To overcome this, the measures of size for households where mos'_h exceeded the sampling interval were capped at the sampling interval (Step a) and the sizes of households where mos'_h was less than the sampling interval were rescaled proportionally such that the sum of mos remained the same for each iteration (Step b). These two steps were carried out iteratively until the converging maximum mos was equal to the sampling interval (at six decimal places).

We were then able to select 4,400 households using probability proportional to size systematic sampling (size equal to the household mos_h). The data set was sorted on PSU, segment, and line before selection, which provided some modest additional stratification. Within each of these 4,400 households, we selected one person with probability proportional to size (mos_i, before iteration began). The iterative procedure optimizes the relative probabilities subject to the constraint.

Received July 28, 2008
Accepted February 9, 2009
Decision Editor: Robert B. Wallace, MD, MSc