Supporting Statement B
for
Population Assessment of
Tobacco and Health (PATH) Study (NIDA)

November 16, 2012

Submitted by:
Kevin P. Conway, Ph.D.
Deputy Director
Division of Epidemiology, Services, and Prevention Research
National Institute on Drug Abuse
6001 Executive Blvd., Room 5185
Rockville, MD 20852
Phone: 301-443-8755

Email: PATHprojectofficer@mail.nih.gov

Table of Contents

LIST OF ATTACHMENTS
(Appearance in Supporting Statement B)

B.   Collections of Information Employing Statistical Methods

The following section focuses on a description of the statistical methods planned for the PATH Study. Section B.1, describes the target population of the PATH Study as well as the respondent universe and the sample composition by various age, tobacco-use, and race-ethnic subgroups. It includes tables summarizing the number of persons in the universe and the expected sample composition. An overview of the sampling frame and sample design is also provided. The section ends with a description of the expected response rates to the PATH Study surveys. Section B.2 describes the procedures for collecting PATH Study data. Weighting and estimation procedures are presented, followed by an elaboration of the degree of precision expected for the analyses of various domains of interest. Section B.3 describes procedures for maximizing the participation and retention of the PATH Study respondents. Section B.4 presents details for the field testing of the PATH Study data collection procedures and operations. Lastly, Section B.5 presents a list of statistical consultants for the PATH Study.

B.1 Respondent Universe and Sampling Methods

B.1a Target Population

The target population of the PATH Study is the civilian household population 18 years of age or older in the U.S. (the 50 states and the District of Columbia), and youth ages 12 to 17. College students will be sampled at their permanent residence rather than at their dormitory as described later in this document. Active-duty members of the military (Army, Navy, Marines, Air Force, and Coast Guard) will be excluded, as will all persons living in group quarters other than college dormitories. The exclusion applies to both institutional and noninstitutional group quarters. Spouses and children of active-duty military living off post in the 50 states and D.C. will be covered.

Consideration was given to sampling other noninstitutional group quarters such as group homes, half-way houses, and shelters. However, important factors weighed against their inclusion: (1) a limited ability to analyze these groups separately given small estimated sample sizes; and (2) the high mobility among persons in such dwellings would lead to high attrition, thereby reducing the information to be gained from this longitudinal cohort study.

B.1b Respondent Universe and Sample Composition

One component of the PATH Study sample design is the selection of a "shadow" sample of 9 to 11 year-olds at baseline (see Section B.1d). Sampled children in this age range are not interviewed until they enter the youth cohort in later waves of the study on reaching 12 years of age. However, for completeness, the estimated respondent universe and sample size of 9 to 11 year-olds are shown in the first row of Table 1.

Estimates of the PATH Study youth respondent universe and expected respondent sample size are shown in the second row of Table 1. Under the planned sample design, the expected number of completed interviews with youth ages 12 to 17 at baseline is approximately 17,070. The estimates in the first two rows of the table are based on data from the 2010 Census and 2010 American Community Survey (ACS).

Estimates of the PATH Study adult respondent universe are also shown in Table 1, which presents the number of persons in specific age, tobacco usage, and race domains derived from population projections. These estimates were formed by applying tobacco prevalence estimates from the (averaged) 2006 and 2007 Tobacco Use Supplements to the Current Population Survey (CPS-TUS) to the adult civilian household population from the 2010 Census, within age/race domains. The table also shows the expected number of survey respondents in the PATH Study baseline sample. Under the planned sample design, the overall expected number of completed adult interviews at baseline is 42,730, including approximately

10,709 young adults (18 to 24 year-olds) and 6,000 Blacks or African Americans (Black/AA)[1].

Sampling rates are also shown for the adult cohort, relative to the 18-24 Black/AA tobacco users domain (which is the most heavily oversampled). Among the many possible definitions of a "tobacco user," the one adopted in this section (including Table 1) mirrors that of the most recent CPS-TUS: a user is anyone who (1) has smoked at least 100 cigarettes in their lifetime and currently smokes cigarettes every day or some days, and/or (2) currently smokes cigars/cigarillos every day or some days, and/or (3) currently smokes a regular tobacco pipe every day or some days, and/or (4) currently uses smokeless tobacco every day or some days.

The sample size estimates in Table 1 are in terms of baseline completed interviews (with or without biological specimens for adults), except that for the shadow sample which represents the number of 9 to 11 year-olds selected at baseline for the purpose of replenishing the youth sample in later waves. The specific subgroups in these tables were selected because they represent the major sampling strata at the person level. Power projections provided later in this submission focus on subgroups of potential analytic interest.

Table 1.   PATH study youth and adult respondent universes and baseline sample sizes

Group	Respondent universe	Baseline sample size	Relative sampling rate
Children 9-11 (shadow sample)	12,639,240	7,728	
Youth 12-17	25,611,322	17,070	
18-24 Black/AA user	835,161	610	1.0
18-24 Black/AA non-user	3,799,471	1,116	2.5
18-24 non-Black/AA user	6,229,213	4,099	1.1
18-24 non-Black/AA non-user	17,610,101	4,884	2.6
25+ Black/AA user	5,155,205	2,051	1.8
25+ Black/AA non-user	19,593,738	2,223	6.4
25+ non-Black/AA user	39,151,562	17,292	1.7

[1] Questions in the PATH study survey instruments that collect data on race or ethnicity will be consistent with the most recent revision of the OMB Statistical Policy Directive No. 15, Race and Ethnic Standards for Federal Statistics and Administrative Reporting. However, the term "Black/AA" as used here refers to anyone who chooses African American or Black as a race category (irrespective of whether one or more race categories are chosen and irrespective of their reported ethnicity).

25+ non-Black/AA non-user	142,189,622	10,455	9.9
Adults	234,564,071	42,730	

B.1c Sampling Frames

The baseline sample for the PATH Study will be selected using a four-stage, stratified probability sample design involving the selection of: (1) primary sampling units (PSUs) consisting of counties or groups of contiguous counties; (2) second-stage sampling units (referred to as segments); (3) mailing addresses; and (4) eligible persons within households occupying dwelling units (DUs) at sampled addresses. In addition to the four stages of selection, a two-phase approach will be used for the fourth stage of sampling (persons within households). The sampling frames to be used at each stage are described here.

For the initial stage of sampling, a PSU frame will be created using the Census 2010 county-level data files. The PSUs will be formed as single counties or groups of contiguous counties, depending on the population size and the end-to-end distance within a PSU. The objective of the PSU formation process will be to simultaneously maximize internal PSU heterogeneity and minimize travel distance within a PSU (e.g., to ensure that the maximum distance is no more than 100 miles), subject to a specified minimum PSU population size of 15,000. Data from the 2010 Census, and data from other sources that will used for stratification purposes, will be appended to the PSU frame. For example, data will be appended from the National Cancer Institute (NCI) small area estimates of county-specific current smoking rates (http://sae.cancer.gov/estimates/tables/both_current.html) and estimates of socio-demographic characteristics from the 5 year ACS.

The second-stage sampling units (referred to as segments) will be based on Census-defined blocks. The frame of segments will be created within the sampled PSUs using the 2010 Census Redistricting Data (P.L. 94-171) Summary File block data, together with address data from the U.S. Postal Service (USPS) Computerized Delivery Sequence Files (CDSFs) of residential addresses. The CDSFs are derived from mailing addresses maintained and updated by the USPS, and they are available from commercial vendors. The

second-stage frame will take data from the most recent CDSF file at the time that the segment sampling is being implemented. Within the sampled PSU, where possible, the associated CDSF addresses will be geocoded to Census blocks and then, as necessary, the blocks will be grouped to create *list segments* of CDSF addresses. Note that post office (PO) box addresses cannot be geocoded and hence will be excluded from this process: thus, DUs that only have a PO box address are not covered by the list segments (however, approximately 90 percent of DUs with PO boxes also have street mailing addresses). Blocks with no population in the 2010 Census will be included in the segment formation process to ensure that all areas are covered. The addresses geocoded to a single block will be used as a list segment if the number of such addresses is larger than a minimum threshold. Otherwise, addresses geocoded to neighboring blocks will be combined to reach the required threshold number of addresses per list segment. Associated with each resulting list segment, will be one or more Census blocks, and the physical boundaries of these blocks will delineate areas of land, referred to as *area segments*. Note that the size of a list segment is based on the number of geocoded CDSF addresses and may well be different from the size of the associated area segment based on 2010 Census data. Differences will arise in part because of date differences but mainly because of geocoding errors made in assigning CDSF addresses to the area segments. Some of the CDSF addresses geocoded to a given area segment may actually be outside the segment's geographical boundaries, and some CDSF addresses that are geocoded to other area segments may be in the given area segment. With the exception of the procedure for providing coverage to addresses not on the CDSFs (discussed later), addresses sampled from a segment will be drawn from the CDSF addresses geocoded to the area segment—that is, from the list segment—irrespective of whether the addresses fall in the area segment or not.

A frame of list segments will be constructed within each sampled PSU by using the prime contractor's "WesBlock" software that is designed to create segments that are contiguous and as compact as possible given the size constraints. The frame of list segments will contain details about the numbers of addresses from the CDSF, the number of households in the associated area segment, and characteristics of the associated area

segments and census tracts from sources such as the 5 year ACS (e.g., urban-rural status, percent Black or African American, percent Hispanic, percent of occupied housing that is owned, and average tract-level household income). In a few rural PSUs, only a small number of geocodeable addresses will appear on the CDSF; in these PSUs, rather than using list segments, conventional area listing procedures will be applied to construct a frame of dwelling units in the sampled segments.

At the third stage of selection, a sample of addresses will be selected from the sampled list segments in the sampled PSUs (except for the few rural PSUs noted earlier). Recent studies indicate that the coverage of the CDSF lists of geocodeable addresses is generally high for urban and large suburban areas, and sometimes reasonably high for parts of rural areas (Montaquila, Hsu, Brick, English and O'Muircheartaigh, 2009; Dohrmann, Han, and Mohadjer, 2007; Iannchionne, Staab, and Redden, 2003; O'Muircheartaigh, Eckman, and Weiss, 2002). To handle any address noncoverage in the CDSF lists, a coverage enhancement procedure, referred to as *address verification*, will be applied for a sample of segments. Although applied only in a sample of segments, this procedure in effect gives coverage for unlisted and non-geocodeable addresses in all segments.

When a segment is selected for address verification, the entire area segment is canvassed by the field interviewer prior to conducting screener interviews at sampled addresses in the list segment, and any addresses not on the CDSF for that list segment are listed for potential inclusion on the supplementary address sampling frame. To handle geocoding errors, the addresses so identified are then matched against the addresses on the CDSF for the ZIP area containing the area segment, and only those not on that CDSF list are retained as a supplementary frame of addresses that will be sampled. The address verification procedure will be applied at higher rates for segments where CDSF undercoverage of geocodeable addresses is likely to be more problematic (e.g., segments where the number of CDSF addresses falls well short of the Census number of households), and the rates of subsampling from the supplementary lists will be determined to counterbalance the segment selection rate for verification. In most urban

areas, the plan will be to sample segments for verification at a low rate and then sample all the addresses on the supplementary frame.

In addition to address verification, a second coverage-improvement procedure, called the *hidden dwelling unit (DU) procedure*, will be applied. The hidden DU procedure is carried out by the field interviewer at the end of the screener interview. Note that a DU is defined as "a group of rooms or a single room occupied as separate living quarters (or if vacant, intended for occupancy as separate living quarters); that is, the occupants do not live with any other person in the structure and there is direct access to the DU from the outside or through a common hall or area." The term "household" includes all persons who occupy a dwelling unit. The hidden DU procedure aims to identify DUs that are attached to the DU where the screener interview is taking place by having the same mailing address or that were not apparent to the canvasser during conventional listing of the segment. Hidden DUs can be part of a unit of a multi-unit building (apartment house), or can be additional or hidden DUs in a single family home (attic or basement apartment or other separate living quarters). Once identified, the hidden DU(s) will be entered into the field interviewer's computer-assisted personal interviewing (CAPI) application, and screening and interviewing will take place within the newly identified unit(s). In cases where a large number of hidden DUs are associated with a sampled address, subsampling of the newly identified hidden DUs will take place.

A special case of the hidden DU procedure applies in cases where the CDSF identifies an address as a drop point, that is, a single address that is recognized as covering a number of DUs. The CDSF includes information on the number of drop units associated with each drop point. Addresses that are drop points with sizable numbers of drop units will be sampled at higher rates than other addresses in order that, in combination with subsampling of the DUs at the drop point if selected, the sampled DUs will retain the desired selection probabilities.

At the fourth stage of selection, the sampling frame for a selected household completing the screener interview will consist of a roster of all the eligible persons in the household. All those 12 years of age and older on the roster

are then eligible to be sampled for either the youth cohort or the adult cohort. In addition, a "shadow sample" of children ages 9 to 11 at the time of screening will be selected for use as a refresher sample for the youth cohort for later waves of the study. After they have turned 12, they will become the refresher sample for the youth cohort.

B.1d Sample Design

As described earlier, the sample will be selected in a four-stage stratified probability design, with a two-phase sample design for sampling the adult cohort at the final stage. The selection processes for these stages are described in turn here.

At the first stage, a stratified sample of 150 PSUs will be selected using probability proportional size (PPS) sampling. The measure of size (MOS) will be defined to be a weighted sum of estimated PSU population counts by the subgroups given in Table 1 where the weights used to construct the MOS are proportional to the expected overall sampling rates to be applied for each subgroup. The PSU population counts by age and race will be obtained from Census Bureau population estimates. The breakdowns of adult age/race groups by tobacco usage will be based on a simple model that takes account of the variability of current smoking rates across PSUs as indicated by the NCI small area estimates. Any PSU that is by itself more than 0.67 percent of the national population (about 2.1 million people) will be treated as a "certainty PSU"; each certainty PSU is in effect a separate stratum. The number of certainty PSUs is anticipated to be about 40.

After accounting for the certainty PSUs, the remaining PSUs will be selected using a carefully stratified design in which the PSUs are selected without replacement and with probability proportionate to size. The stratification factors will include such variables as the geographic region, NCI estimates of county-specific smoking rates, MSA status, percent minority population, poverty rate, and other variables where appropriate. Around 55 approximately equal sized strata (in terms of aggregate MOS) are expected to be formed.

Within the selected PSUs, segments will be formed, and a systematic PPS sample of about 40 segments will be drawn within each noncertainty PSU, more in most certainty PSUs, for a total of 6,000 segments. The systematic selection will be with respect to a sort of the segments. The sort variables to be considered include urban-rural status, percent of occupied housing that is owned, race/ethnicity, and possibly average tract-level household income (based on data from the American Community Survey) for the associated area segment.

At the third stage of sampling, a systematic sample of addresses will be drawn from the CDSF list frame for the list segments, or from the conventional list frame constructed for the area segments for the PSUs for which the proportion of geocodeable CDSF addresses is very low. For segments in which the verification procedure is applied, a sample of any supplementary addresses will also be selected. The hidden DU procedure will be applied at all sampled addresses, with the end product being a sample of households.

A roster of all the members of each sampled household will then be constructed by interviewing a household informant, together with information on the person's age and, for adults, on their race (Black/African-American vs. all others), and tobacco use. The three components of the sampling of household members are as follows:

1. Shadow sample

   If any children ages 9 to 11 are in a household, one is selected at random for the shadow sample. Sampled children in this age range may enter the youth cohort in later waves of the study on reaching 12 years of age.

2. Youth cohort

   If any youth ages 12 to 17 are in a household, one or two will be selected at random for the youth cohort.

3. Adult cohort

   No more than two adults will be sampled for the adult cohort.

Given a special analytic interest in monozygotic and dizygotic twins in the youth cohort, the shadow and youth cohort sampling procedures are modified when households containing twins are encountered. In such households, a twin pair will be sampled, and if another youth or youths are in the same sample group (shadow or youth) as the twins, those youth will be given a probability of 1/3 of also being selected with at most one being selected.

The sampling method for selecting adults for the first phase of the adult cohort depends on the desired selection probabilities for each of the adult subgroups listed in Table 1 within the sampled PSU. To describe the method, let p_l denote the desired within-household sampling rate for an adult *l*. This probability depends on the person's age, race, and tobacco use. Let $P_k = \Sigma_{l \in k}\, p_k$ be the sum of these probabilities for all adults in the household. The following two classes of households can be distinguished:

1. Households with $P_k \leq 2$. Select $0, 1$ or 2 adults by a systematic PPS sample with measures of size p_l and an interval of 1.

2. Households with $P_k > 2$. Select two adults by PPS with an interval of 1 and with adjusted probabilities $p_l(2/P_k)$.

The second phase of sampling is included to address classification errors in the responses of household screener respondents, in particular the misclassification of a sampled person as a non-tobacco user when the self-report would indicate the person is a user. At the first phase, the sampling rates for non-users are kept within reasonable bounds, compared to the rates for users, in order to ensure that the weights of any persons sampled at the first phase as non-users, who then report themselves at the second phase to be users, are not too much larger than the weights of those who are correctly classified as users at the first phase. Misclassification in the other direction — with the household informant reporting the person as a user when the person then reports him- or her-self as a non-user — will be handled by deselecting some members of this group so that those retained have the same sampling rates as other non-users. All the sampled persons classified by self-report as users are retained at the second phase. Those

correctly classified as non-users at the first phase are subsampled at the second phase to achieve the desired sample sizes for non-users.

B.1e Group Quarters

The types of noninstitutionalized group quarters that are of interest to the PATH study include only college dormitories. College students living in dormitories and fraternity and sorority houses will be sampled through their permanent residence. If a student who lives in a dormitory for much of the year is identified as a sample person and is at home (their permanent residence) when the screening occurs, an attempt to administer the interview will be made before the student returns to the dormitory. Otherwise, a time will be found during the field period when the student will be at home and an interview will be scheduled for that time. If this is not possible, the student will be contacted and interviewed on campus if the dormitory is close to any sampled PSU (which need not be the PSU of the family residence). Identifying students in dormitories via their family residence is a simpler process than constructing a separate dormitory sampling frame from which to select students. It avoids the costs and complications of contacting and gaining permission from college and university officials, of obtaining and sampling from lists of dormitories, and of listing and sampling within selected dormitories.

B.1f Expected Response Rates

As noted earlier, once a dwelling unit is selected, a household screener will be administered in the field to determine the race, age, and tobacco usage of each adult as well as the age of each child in the household. The expected response rate for this procedure is 87 percent. Section B.3 describes the various approaches the PATH Study will employ to achieve this target. This response rate is based on the similar rate achieved by the 2011 National Survey of Drug Use and Health (NSDUH), conducted by the Substance Abuse and Mental Health Services Administration (SAMHSA). (http://www.samhsa.gov/data/NSDUH/2k11Results/NSDUHresults2011.htm,

retrieved November 6, 2012). NSDUH does not offer any incentive to the screener respondent. Because the burden for the screener respondent (and for the household as a whole) will, on average, be higher for the PATH Study than for the NSDUH sample, the PATH field test includes an experiment to determine whether the PATH Study can achieve a comparable screener response rate without a screener incentive and, if not, whether the nominal monetary amounts of $5 and $10 will help to facilitate the desired screener response rate. This incentive experiment is described further in Attachment 19.

In terms of screening, the expected eligibility rate for households will be close to 100 percent because a negligible number of households in the U.S. are comprised solely of military personnel on active duty, children who are 11 years or younger, and persons who are unable to conduct an interview in English, Spanish, Mandarin, Cantonese, Korean, or Vietnamese. Depending on the age, race, and tobacco usage of the adult persons in the household *as reported by the household informant*, up to two adults will be selected at the first phase of sampling. The selected adults will proceed to the second phase of sampling where they will be administered a short series of questions at the beginning of the extended interview to determine their self-reported tobacco usage. Based on these self-reports it is expected that approximately 68 percent of the adults selected at the first phase of sampling will be retained at the second phase of sampling to be administered the full extended interview. At the same time, up to two youth ages 12 to 17 can be sampled from the household (or in the case of multiple births, up to 3 youth per household). Within each household, independent sampling will be conducted for adults and for youth. For both sampled adults and sampled youth, the response rates for the extended interviews are expected to be 90 percent for the baseline, 92 percent for Wave 2, 95 percent for Wave 3, and 96 percent for each of the remaining follow-up waves. These attrition rates are based on 2008-2011 Medical Expenditure Panel Survey (MEPS). A series of measures will be taken to secure these response rates, as described in Section B.3. Thus, the overall baseline response rate is expected to be 78 percent for both adults and youth (i.e., the product of the expected screener response rate and the expected person-level response rate). Table 3
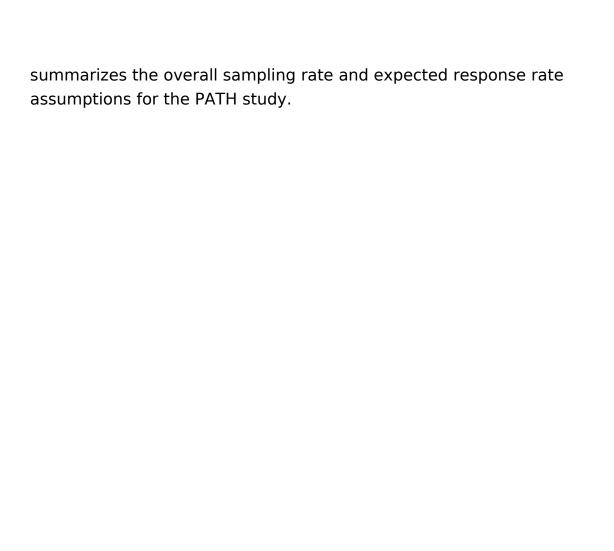
summarizes the overall sampling rate and expected response rate assumptions for the PATH study.

Table 2.   Overall sampling rate and expected response rate assumptions for the PATH study at baseline

Sampling unit	Assumed rate	Expected number
Primary sampling unit (PSU)	---	150
Area segments/CDSF segments	40 per PSU	6,000
Addresses	22.1 per segment	132,668
Occupied dwelling units	88.6%	117,544
Households completing screener enumeration	87%	102,263
Adult sample (persons ages 18+)	100%*	102,263
Eligible households with adults		
Number of adults sampled at first stage	Up to 2 per HH	70,000
Number of adults completing second-phase sampling questions at beginning of extended interview	90%	63,000
Number of adults retained at second phase of sampling and completing full extended interview	68%	42,730
Number of adults completing extended interview who provide buccal cells	85%	36,321
Number of adults completing extended interview who provide urine	85%	36,321
Number of adults completing extended interview who provide blood	65%	27,775
Number of adults completing extended interview who provide all biospecimens	65%	27,775
Youth sample (persons ages 12-17)		
Eligible households with youth	16%	16,362
Eligible households reporting youth	90%	14,726
Number of sampled youth	Up to 2 per HH	18,967
Number of youth completing extended interview	90%	17,070

* A very small number of screened households may contain only persons under 18 or on active duty.

B.2 Procedures for the Collection of Information

B.2a Overview

The PATH Study involves four main components. These components are: (1) an automated CAPI (Computer-Assisted Personal Interviewing) household screening instrument, (2) automated ACASI (Automated Computer-Assisted Self-Interview) extended instruments (separate instruments for youth and adults), (3) an automated CAPI parent instrument, and (4) collection of

biospecimens from adults (buccal cells and urine at baseline and each follow-up wave, and whole blood at baseline and in the second follow-up wave). All of the main components, except the screening instrument, will be repeated in three follow-up waves. Collection of biospecimens is not a requirement for PATH Study participation; however completion of an extended interview at the first home visit is required.

The primary objective of the field interviewer working on the PATH Study is to obtain complete and accurate information from sampled persons in each eligible household in their assignment. This requires that the field interviewer have a thorough understanding of the survey's protocol, as well as an understanding of the techniques required to gain the respondent's cooperation and maintain rapport through the interaction. All field interviewers working on the PATH Study will receive extensive in-person training on the exact procedures to be followed in the administration of the data collection instruments themselves, as well as techniques to gain cooperation, such as understanding the importance of the study, answering respondent questions, and addressing respondent concerns.

The training provided to field interviewers will be in two forms: home study and in-person. The 11-hour home study program will be designed to introduce trainees to the PATH Study, with a focus on the respondent contact materials. The home study will also provide field interviewers with practice in gaining cooperation and establishing rapport. In-person training techniques are designed to maximize trainee involvement, maintain the interest of the trainees, and produce well-trained field interviewers who have the necessary skills for gaining respondent cooperation, correctly answering questions about the study, and adeptly completing all components of the interviews. Training materials will be developed by experienced PATH Study team members. In the 4-day in-person session, field interviewers will be trained on techniques for obtaining consent; conducting the CAPI screener, ACASI extended interviews, and CAPI parent interview; collecting buccal cell and urine samples; issuing respondent incentives; and completing administrative procedures such as data transmission and reporting to the supervisor. In addition, experienced phlebotomists will be trained to return to the homes of consenting adults to collect whole blood samples.

In addition to the in-person training, field interviewers will be provided with a field interviewer manual, providing detailed reference materials on locating sampled addresses, determining household membership, the interviewing process, questionnaire content, and biospecimen procedures. The phlebotomists will be provided with a study-specific phlebotomist manual on collecting whole blood.

During the data collection period, numerous quality control procedures will be used to ensure that field interviewers are following the specified procedures and protocols and that the data collected are of the highest quality. Field interviewers who successfully completed training, but show any area of potential weakness, will be observed in-person at least one time by a supervisor or home office staff members. Observing field interviewers conducting their job in the field is a very effective way of monitoring their skills to conduct the interview, as well as their adherence to survey procedures. It also provides the observer with an appreciation of the field interviewers' tasks and provides the opportunity to experience first-hand the administration of the PATH Study instruments and biospecimen collection procedures. Observations will be concentrated in the early weeks of data collection so that problems are detected as early as possible, to provide corrective feedback to the field interviewers.

Brief quality control interviews will be conducted to verify that an interview was administered or attempted as reported by the field interviewer. Quality control procedures will be implemented to verify at least ten percent of each field interviewer's finalized work to ensure that the interview was conducted according to study procedures. This includes cases finalized as complete, as well as those with non-complete dispositions, such as vacant or refusal. Quality control will begin early in the data collection period to allow for any identified problems to be addressed immediately. Quality control interviews will be conducted by separate trained data collection staff over the telephone, whenever possible. However, if unable to complete a quality control interview via telephone (e.g., the dwelling unit is vacant), the interview will be assigned to an experienced, specially trained field interviewer who will conduct the quality control procedure in person.

Additionally, throughout the field period, supervisors will remain in close contact with the field interviewers. Scheduled weekly telephone conferences will be held in which all non-finalized cases assigned to the field interviewer will be reviewed to determine the best approach for working the cases and the need for additional resources.

Management staff at all levels will have access to a supervisor management system, including automated management and production reports that will be used to monitor the data collection effort and ensure that the data collection and quality control goals are being attained. Field interviewers will be required to transmit data on a daily basis. Data will be transmitted to a secure server at the office of the prime contractor, which will then be used to update the automated management reports. These data are also used to produce weekly reports that might provide evidence of suspicious field interviewer behavior, such as overall interview administration length, individual instrument administration time, amount of time between interviews, interviews conducted very early in the morning or late in the evening, and number of interviews conducted per day.

B.2b Household Screener

The random selection of up to two adults and two youth (unless a household includes twins, in which case additional youth could be selected) per eligible household (as described in Section B.1) is conducted through the use of an automated screening instrument (see Attachment 2). The screener uses a full household enumeration process to collect information on age for each reported household member, and race, active military service status, and tobacco use for each adult household member. The relationship of all household members to the screener respondent also is collected. The screener respondent will be an adult household member age 18 or older. In addition to household enumeration information, household and each sample person's telephone numbers are collected to allow the recontact of the household for quality control purposes, or to set appointments for the extended and parent interviews if the sample person is unavailable at the

time of the screening. Finally, if the mailing address differs from the street address, the household mailing address is collected. Mailing address allows written follow-up with nonresponse cases and regular contact with respondents between data and biospecimen collection waves, as discussed in Section B3.

The proposed sampling algorithm for selecting up to two adults and two youth (except in the case of twins) per household has been programmed within the CAPI screener software. To check that the screener is working properly, it will be tested extensively by professional software testers.

B.2c Extended Interview

The data collection procedures differ for adults and youth.

Adults

Following the administration of the screener, if the selected sample adult is available and has an adequate amount of time to complete the interview, the field interviewer: (1) obtains informed consent (see Attachment 11); (2) administers the adult extended interview, which includes gathering additional contact information about the adult; (3) obtains consent for the biospecimen collection; (4) gathers the biospecimens (typically only a buccal cell sample); (5) arranges a follow-up appointment for a phlebotomist to collect a blood sample and typically a urine sample; and (6) pays the incentive to the respondent at the completion of the first home visit. (The biospecimen collection is discussed further in Section B.2d.) If a sample adult is unavailable or unable to complete the interview at that time, the field interviewer will attempt to schedule an appointment for a return visit or, at a minimum, determine the best time for a return visit.

After obtaining consent, the field interviewer provides a brief automated tutorial on using ACASI and launches the automated ACASI extended interview. The first part of the extended interview is the individual screener;

these items may confirm or contradict the information provided in the first-phase household screener by the screener respondent. Depending on the individual's self-reports (e.g., on tobacco usage), the sample person may be de-selected and not asked to complete the remainder of the extended instrument. As required throughout the interview, the field interviewer will aid the sample person in providing a response. At the end of the extended interview, the field interviewer gathers additional contact information for that person, and asks the respondent to consent to providing biospecimens. (See Section B.2d.)

The sample adult who completes the extended interview or is excluded based on the his/her responses to the individual screener items (which constitutes the second phase of screening described in Section B.1d) will receive $35 (the extended interview incentive) as a thank you for his/her time to complete the interview. These respondents also will receive a thank you letter (Attachment 7). A refusal conversion letter will be sent to sample adults who initially decline to participate or are difficult to contact (Attachment 17).

Youth

Following the administration of the screener, if the parent or guardian of the selected youth is available and has an adequate amount of time, the field interviewer: (1) obtains parent permission for the youth to participate; (2) obtains consent for the short parent interview; and (3) administers the CAPI parent interview, which includes gathering additional contact information about the youth from the parent. If a parent of a sampled youth is unavailable or unable to participate at that time, the field interviewer will attempt to schedule an appointment for a return visit or, at a minimum, determine the best time for a return visit. The youth interview will not be conducted until parental informed consent has been obtained. The parent who completes a parent interview for the youth will receive $10 as a thank you for his/her time to complete the interview.

For a selected youth with parental permission, if the youth is available and has an adequate amount of time to complete the interview, the field interviewer obtains youth assent (see Attachment 11) and then attempts to complete the automated ACASI extended instrument. If a sample youth is unavailable or unable to complete the interview at that time, the field interviewer will attempt to schedule an appointment for a return visit or, at a minimum, determine the best time for a return visit.

After obtaining assent from the selected youth, the field interviewer provides a brief automated tutorial on using ACASI and launches the automated ACASI extended interview. As required throughout the interview, the field interviewer will aid the sample person in providing a response.

The youth respondent who completes the extended interview will receive $25 (the youth extended incentive) as a thank you for his/her time completing the interview. The parents of youth respondents will receive a thank you letter (Attachment 7). A refusal conversion letter will also be sent to the parents of respondents who are difficult to contact (Attachment 17).

Burden Reduction by Avoiding Redundant Data Collection

When feasible and desirable, the PATH Study will avoid collecting data that are already available from a previously collected source. For example, the parent interview collects personal information about the parent of a sampled youth, some general characteristics of the household as a whole, and information about the youth, plus contact information to support reaching the parent and youth for future data collection activities. Since more than one youth may be sampled per household, one parent may be asked to respond to a parent interview in regard to more than one youth. In this instance, the parent will not be asked to again provide his or her personal information, the household information, or the contact information after the first instance of the parent interview. Only the information relevant to each sampled youth about whom the parent is providing information will be collected after the first administration of the parent interview.

There are a few instances where the PATH Study will purposely collect some information that has been previously provided to validate previous information, to give respondents the opportunity to consider their answers in a private setting, and to collect information that provides broader context and background to the respondent on particular items. The main instance where this occurs is in the second-phase adult individual screener.

Among other purposes, the household screener collects a minimum amount of high-level information about each adult's tobacco use in order to classify him/her sufficiently for potential selection to the study based on the PATH Study sampling algorithm..

The first-phase household screener obtains tobacco use information about all adults from the single household respondent. Various reasons why this approach may yield inconsistent or imperfect information are described in Section A.2b. To obtain more complete, consistent information from an individual adult, the second-phase screener (i.e., the adult individual screener) is used to ask a more extensive panel of tobacco use questions. Using a second-phase screener such as this helps to reduce bias and increase the validity of responses obtained from an individual respondent rather than from the household respondent who completed the first-phase screener. Even if the person who completed the first-phase screener is the same individual who completes the second-phase screener, the second-phase screener is designed to reduce bias and enhance the validity of responses because: (1) questions are asked in a more private setting using ACASI (rather than CAPI); (2) questions are more detailed and given a more detailed context; and (3) questions are asked in an open format such that it is clear to the respondent that there are no "right" or "wrong" answers.

B.2d Biospecimens

The field interviewer will ask an adult who completes an extended interview to consent to provide biospecimens as part of the PATH Study. However, providing biospecimens is voluntary and not a condition of participation.

Completion of the extended interview at the first home visit is required from all respondents.

Buccal Cells and Urine

The biospecimen collection activities depend on whether the adult respondent consents to provide a blood specimen, which would occur at a second home visit by a health professional/phlebotomist. If the adult consents, both a urine and blood sample will be collected at the follow-up visit. For adults who consent to provide buccal cells, the field interviewer will collect those specimens following the completion of the interview. If the adult declines consent to provide a blood specimen, but agrees to provide a urine specimen, the field interviewer will collect both the buccal cell and urine specimens following the interview. The field interviewer will enter information on the respondent's recent use of tobacco products into the laptop computer (see Attachment 2), and provide written and oral instructions to the respondent for collection of the buccal cells (and urine specimen, if applicable). The field interviewer will pack the specimen(s) and ship them to the PATH Study biorepository.

The respondent who provides a biospecimen during a first home visit will receive $10 as a thank you for the time involved to provide the sample.

Blood

For adults who consent to provide blood, the field interviewer will administer blood suitability exclusion questions for blood collection (see Attachment 2). If the respondent has no suitability exclusions, the field interviewer will schedule the appointment for the visit by the phlebotomist to obtain the blood specimen (and the urine specimen, if applicable). After the initial home visit by the field interviewer, the phlebotomist will contact the adult to confirm the appointment for collecting a blood (and urine) specimen.

Upon visiting the respondent's home, the phlebotomist will re-administer the blood suitability exclusion questions (see Attachment 17) for blood collection and request that respondents answer items about his/her recent use of tobacco products onto a paper form (see Attachment 17). The phlebotomist will provide written and oral instructions to the respondent for collection of the blood (and urine) specimen(s). The field interviewer will pack the specimen(s) and ship them to the PATH Study biorepository.

The respondent who provides a blood specimen (or both a urine and blood specimen) during a second home visit will receive $25 as a thank you.

B.2e Weighting and Estimation Procedures

Sample weights will be developed for the PATH Study respondents to permit estimation for and inference about the population from which the sample is drawn. The sample weights will be produced to accomplish the following objectives:

1. Permit the appropriate development of estimates, taking account of the fact that not all persons in the target population will have the same probability of selection;

2. Limit the potential for biases arising from differences between cooperating and noncooperating sample persons and households;

3. Use auxiliary data on known population characteristics in such a way as to reduce coverage biases and benchmark survey estimates to the corresponding population totals;

4. Reduce the variation of the weights and prevent a small number of observations from dominating domain estimates; and

5. Facilitate sampling error estimation appropriate to the complex sample design.

The data used in weighting will undergo careful edit, frequency, and consistency checks to prevent errors in the sample weights. The checks will be performed on items to be used in the weighting procedures and will be

limited to records that require weights. These checks are important because errors in the weights can have sizable effects on the survey estimates.

The first step in the weighting process is to compute the selection probabilities for all households sampled (responding households and nonresponding households). The household base weights are then the inverses of these selection probabilities. The household base weights of responding households (i.e., those for which the screener is completed) are then inflated to compensate for the nonresponding households. The adjusted household weights are then the starting point for the computation of the person weights.

Persons are selected with different probabilities within responding households in order to achieve required sample sizes by tobacco use, age, and race. The next step is then to modify the adjusted household weights to create person base weights that compensate for the unequal selection probabilities of sampled persons (respondents and nonrespondents). At this point, a decision will need to be made as to what constitutes a "response." Persons who start the interview but break off early on are commonly treated as nonrespondents.

More significant for the PATH Study is the response classification of those adults who complete the interview but do not provide any of the biospecimens, and of those who complete the interview and provide the buccal cell and urine samples but not the blood samples. A complication here is that some of the biospecimens will turn out not to be analyzable, so they are nonresponses. However, because the biospecimens will not be analyzed until later, the biospecimens unable to be analyzed will not have been identified when the weighting is being conducted.

Two alternative approaches can be used for handling component nonresponse (here, biospecimen nonresponse). One is to treat the component nonresponse as a set of item nonresponses in a respondent record, using imputation as the means of compensation for the missing data. In this case, the analytic data file for the baseline data collection would comprise all sampled adults who completed the interview, irrespective of

whether they provided any of the biospecimens, and all sampled youth who completed the interview.

The other approach for handling component nonresponse is to create separate sets of weights according to which components were completed. For example, for the PATH Study, one set of weights could be for all the adults who completed the interview and these weights would be used for analyses that are confined to the interview data. A second set of weights could be computed for all adults completed the interview and the buccal cell and urine samples, for use in analyses that required those biospecimens only. A third set of weights could be computed for adults who completed the interview and provided all three biospecimens. Such an approach maximizes the sample size for each type of analysis. However, computing all sets of weights may not be worthwhile if the sample size for a data set that is more inclusive than the data used in a given analysis is not much smaller than the data set that contains only the data required.

Given the delay in the analyses of the biospecimens, no immediate decision will be made between these two approaches. At this point, only one set of adult weights will be constructed : the weights for all those who complete the interview. Decisions about whether to impute for missing biospecimens or whether to create separate sets of weights for different patterns of biospecimen response can be deferred until the biospecimen data are to be analyzed. This approach also allows for adults who may refuse to provide one or more biospecimens at baseline, but agree to do so at a subsequent wave of the study.

The steps described in detail in the next section indicate the weighting process to be used for the development of the baseline weights for the respondents to the baseline interviews. For subsequent waves, nonresponse adjustments that account for cohort attrition across waves will be undertaken to produce longitudinal weights. In addition, sampled persons who age into the youth or adult cohort study (i.e., reach age 12 or 18) in subsequent waves will need to be assigned weights for cross-sectional analyses for the wave they enter, and for cross-sectional and possibly also for some longitudinal analyses thereafter.

Development of the Sample Weights for Baseline Respondents

Screener Base Weights

The first step in the development of the baseline person weights is to calculate base weights for all sampled households. The screener base weight initially will be computed as the reciprocal of the product of the household's selection probability. The final, adjusted, screener weight will include the adjustments needed to reflect the selection of nonresponding households.

The screener base weight is given by

$$W_{ijk}^{base,SCR} = 1/P_{ijk}$$

where P_{ijk} represents the overall probability of selection of household *k* in segment *j* of PSU *i.* For most cases, households will be sampled straightforwardly from the USPS address list, in which case P_{ijk} is simply the product of the PSU, the segment-within-PSU, and the address-within-segment probabilities. The same probability also applies to households sampled though the address verification or hidden DU procedures provided all households "discovered" at the sampled address are sampled. If a sample of households is taken at the address, then the probability of sampling the household given the address has to be added as an additional multiplier. The value of P_{ijk} for households sampled from the address verification procedure is the product of the PSU and segment-within-PSU probabilities multiplied by the probability of the address verification procedure being applied and the probability of the household being selected given that the address verification procedure was used.

Adjustment for Nonresponse to the Screener

The household base weights are computed for all sampled households. However, the screener will not be completed for all of them. Adjustments will

therefore be made to the base weights of responding households to compensate for the nonresponding households. All adjustments will be made within weighting classes based on information available for both responding and nonresponding households, namely the segments in which they are located. Census 2010 data at the area segment level and geographical proximity will be used to group segments into weighting classes.

Then, within a weighting class, the base weights for the responding households will be inflated proportionately so that they produce the same sum as the sum of the base weights of the responding and nonresponding households combined.

Person Base Weights

To produce unbiased estimates, a weighting factor is needed to account for the within-household selection rate. The person base weights will be computed as the product of the screener nonresponse-adjusted weight and the reciprocal of the within-household probability of selection for person *l* within household *k* of PSU *i* and segment *j*, as shown in the following formula:

$$W_{ijkl}^{base,BQ} = W_{ijkl}^{nr,SCR} / CP_{ijkl} \text{ ;}$$

where

CP_{ijkl} = the within-household probability of person *l* being selected into the sample, which will depend on the number of persons in household *k*, their ages, races, and tobacco use, and

$W_{ijkl}^{nr,SCR}$ = the screener nonresponse-adjusted weight.

Adjustment for Person-Level Nonresponse

Similar to the adjustment for screener nonresponse, a nonresponse adjustment will be performed to account for nonrespondents to the extended

interview. The weights of respondents to the extended interview will be inflated to account for the nonrespondents. In addition to segment identification available for the screener nonresponse adjustment, screener variables such as age, gender, race/ethnicity, and tobacco usage also can be used to form weighting classes. A variety of methods, such as CHAID (Chi-squared Automatic Interaction Detector), logistic modeling of response propensity, and data mining, exists for determining the weighting classes.

Trimming

Trimming is a process in which inordinately large weights are reduced. It is used because very large weights can substantially increase sampling errors. A trimming algorithm will be used to reduce the variation in the nonresponse-adjusted weights. In general, trimming procedures introduce some bias into the sample estimates. However, when the trimming adjustment is applied to only a very small number of weights, the expectation is that the reduction in the sampling error component of the overall mean square error will more than offset the increase in bias. A preassigned rule will be applied within prespecified sampling and analytical domains to determine which weights should be trimmed.

Computing Final Weights—Poststratification Through Raking

Undercoverage of the target population is a common problem in surveys. Undercoverage occurs when some population units are not included in the sampling frame and have no chance of being selected into the sample. It also can arise in household enumeration when not all the eligible household members are enumerated for sampling. Techniques such as the address verification and hidden DU procedures are used to improve coverage rates for households. Methodologically sound approaches to screening households can limit the degree of undercoverage experienced during household enumeration. Nevertheless, all surveys are subject to some amount of undercoverage, and the PATH Study will be no exception. To account for undercoverage and other sources of bias remaining after the nonresponse

adjustment, the person weights resulting from the previously applied steps will be adjusted to independent estimates of population parameters. This will be accomplished by "raking" (as described here) the weights so that numerous totals calculated with the resulting full sample weights will agree with population totals from the Census Bureau's Population Estimates Program and/or the American Community Survey (ACS).

A particular form of poststratification referred to as raking ratio adjustments is planned. The final sampling weights will be computed by raking the weights to known population totals. In poststratification, classes are formed from cross-tabulations of certain variables. In some instances, such cross-tabulations may lead to sparse cells, or population distributions may be known for the marginal but not the joint distributions of variables used to define the weighting classes. Weighting class adjustments based on small cell sizes can result in a large amount of variation in the adjusted weights. Raking ratio adjustments are useful for maintaining the weighted marginal distributions of variables used to define weighting classes. For this type of adjustment, population distributions are required for the marginal distributions of the weighting class variables and not for their joint distribution.

The weights of all persons who complete the interview will be included in the raking. Segment-level and screener variables can be used to form raking cells, as well as variables from the extended interview.

Specially developed SAS macros will be used to compute the weights for the PATH Study sample. These macros perform such tasks as cell weighting adjustments for nonresponse, poststratification, raking, generalized regression estimation, creation of replicates for variance estimation, and weighting adjustments (i.e., nonresponse adjustment, poststratification, generalized regression estimation, and raking) of the replicate weights.

Replicated Weights for Variance Estimation

Variance estimation must take into account the sample design. In particular, the estimate of sampling variance for any statistic should account for the effects of clustering, stratification, unequal selection probabilities, and the use of nonresponse, trimming, and poststratification adjustments. For the PATH Study, treating the data as having been selected by simple random sampling will produce underestimates of the true sampling variability.

Several alternative replication methods have been developed for computing valid variance estimates for survey estimates based on complex sample designs. The plan for the PATH Study is to use the jackknife method. It can be used to estimate variances for most statistics. The jackknife method drops one PSU from a stratum and increases the base weights of the units in the other PSUs in the stratum to compensate. An estimate of the statistic of interest, say θ, is then computed from the reduced sample. This process is repeated, dropping PSUs in turn to create a series of estimates of θ, say $\hat{\theta}_g$ for replicate g. A general version of the jackknife drops each of the PSUs in turn. For this version, the variance of the estimate of θ based on the full sample, $\hat{\theta}$, is computed as

$$v(\hat{\theta}) = \sum_{g=1}^{G} \frac{(n_h - 1)}{n_h}(\hat{\theta}_g - \hat{\theta})^2$$

where n_h sampled PSUs are in stratum h and $G = \Sigma n_h$.
After computing the base weights for each replicate, all the weight adjustment steps leading to the final person weight will be performed on each replicate. By repeating these adjustments on the revised base weights for each replicate, the impact of these procedures on the sampling variance of the estimator $\hat{\theta}$ is appropriately reflected in the variance estimator $v(\hat{\theta})$.

Various modifications may be made if the number of replicates with the general procedure is very large. In the case of the PATH Study, this issue will arise only with respect to the certainty PSUs (in reality, they are strata). For these PSUs, the actual PSUs are in fact the segments. A standard procedure of combining groups of segments will be applied to avoid an excessive

number of replicates. This combining procedure does not lead to any bias in the variance estimation.

A number of programs can be used for computing variances with replication methods including the prime contractor's "WesVar" software, which is freely downloadable from the web. Alternatively, strata and PSU identifiers enable a linearization approach to variance estimation to be used, if required.

Longitudinal Weighting

The previous discussion describes how the weighting will be carried out for respondents to the baseline wave. At the second wave of the cohort study, interview data will not be obtained for some of the baseline respondents. Some form of compensation is required for the resultant missing data. Those who respond at both waves will constitute the data set for longitudinal analyses. For cross-sectional analysis, those sampled 17 year-olds who have reached the age of 18 in the intervening year will be added to the adult cohort representing the adult population of inference.

Two alternative approaches can be used for compensating for baseline respondents who do not respond at the second wave: imputation and weighting adjustments (see, for example, Kalton, 1986). Each approach has its advantages and disadvantages – a recommendation for using one of them will be made after the data are collected and patterns of nonresponse can be assessed.

The imputation approach keeps the second wave nonrespondents in the analytic file, imputing all their missing second wave responses based on their baseline data. Performing all these imputations in an effective way that does not distort relationships between items both cross-sectionally and longitudinally is the major concern with this approach. Until recently, this concern has resulted in the weighting approach generally being preferred. However, recent developments in imputation software, such as the prime contractor's "AutoImpute" software, make the imputation approach more

competitive. With this approach, the baseline weights of interview respondents are not altered for longitudinal analyses.

To date, the usual way to compensate for wave nonresponse has been by a weighting adjustment. Because there is so much information about the second wave nonrespondents will be available from their baseline responses, the challenges with this approach are to select the auxiliary variables to be used in making the adjustments and to determine the form of adjustment to use. For example, Rizzo, Kalton, and Brick (1996) describe analyses they performed under a contract with the Census Bureau to examine these issues for handing panel attrition in the Survey of Income and Program Participation.

A complication arises in later waves if a respondent misses one wave but returns to the cohort in the following wave. With the imputation approach, the imputed values for the missing wave should be made consistent with the responses for the adjacent waves. With the weighting approach, those missing a wave can be incorporated in cross-sectional estimates for the later wave, but they will not provide data for longitudinal analyses involving the missing wave. A possible compromise approach is to apply weighting adjustments for second-wave respondents for analyses at that time, but then to impute responses for those nonrespondents at the second wave who respond at the third wave, incorporating both first and third-wave responses in the imputation model. NIDA and FDA will work with the prime contractor to determine the best approach to use in the analyses for handling respondents with a missing wave of data that is bounded by reported data for adjacent waves.

Those 9 to 11 year-olds selected as part of the shadow sample will be included in the baseline weighting process. Their weights will be the household base weights adjusted for nonresponse at the household level. These weights will serve as the "base weights" for the shadow sample members when they become 12 years old and join the youth cohort. Consideration may be given to doing a poststratification adjustment for 12 year-olds each year to help ensure that they are fully reflected among the group ages 12 to 17.

B.2f Expected Levels of Precision of the PATH Study

Many of the major objectives of the PATH Study require the estimation of the
prevalence of various tobacco-related attitudes, behaviors, and health
consequences, and changes in these measures over time. In order to achieve
these objectives, the PATH Study was designed to produce reliable estimates
for these characteristics for various population subgroups at the first or
baseline wave and over follow-up waves. The characteristics of most interest
are dichotomous, having "yes" or "no" outcomes. The percentage of "yes"
responses is denoted by p and represents the prevalence rate of a particular
characteristic (e.g., cigarette smoking). Based on previous research and the
accumulation of professional experience, the majority of characteristics
measured in the PATH Study are expected to have magnitudes of prevalence
exceeding 10 percent, while the expected magnitude of a few characteristics
(such as initiation of tobacco use) will lie between 1 and 5 percent.

One measure of the precision associated with cross-sectional prevalence
rates is the relative standard error (RSE), which is defined as the standard
error divided by the prevalence estimate, expressed as a percentage. More
specifically, $RSE(p) = 100 * Standard\,Error(p) / p$, where the standard error is given
by the square root of the variance of the estimate, taking into account the
complex sample design of the PATH Study. A measure of power associated
with longitudinal analyses of change in prevalence rates is the minimum
detectable absolute difference (MDAD). Herein, the MDADs represent the
smallest change (up or down) from a given baseline prevalence rate that can
be detected with 80 percent power using a two-sided test at the 5 percent
level of significance, taking into account the complex sample design of the
study. The impact of the various complex features of the sample design on
variances, and therefore on RSEs and MDADs, is reflected through inflation
factors called design effects (DEFFs). The extent to which these design
effects exceed one indicates the extent to which the variance of an estimate
based on the complex sample design is greater than the corresponding
variance based on a simple random sample (SRS) design. Several key

features to the PATH Study sampling design contribute to the overall design effect.

The first feature is the clustering at both the PSU and segment levels. In general, for a fixed sample size, the greater the number of units to be sampled per cluster, and the more homogeneous the sampling units are with respect to a characteristic of interest within clusters, the greater the DEFF and hence the inflation in the variance (resulting in decreased precision). The level of homogeneity within a cluster is reflected through two types of intraclass correlations: ρ_1 for PSUs and ρ_2 for segments. Note that ρ_1 and ρ_2 will vary in value for different characteristics of interest. The expected standard errors for prevalence estimates for the PATH Study have been calculated taking into account the contributions due to clustering at both the PSU and segment levels under the assumptions that the intraclass correlations (ρ_1, ρ_2) are (.01, .05). These values were based on estimates taken from various sources in the survey literature. The calculations reflect the fact that "certainty PSUs" are, in fact, strata not PSUs, so that no contribution to the variance from clustering at the PSU level occurs for these PSUs. With 150 PSUs selected, approximately 40 certainties, representing 35 percent of the U.S. population, are estimated.

A second feature of the PATH Study design that contributes to the overall sampling variability is the plan to sample adults with different selection probabilities according to their age, race, and tobacco use (the latter both as reported by the household screener respondent and as self-reported by the adult at the second phase of screening). The unequal weighting DEFFs due to this feature of the sample design are expected to range from 1.00 to 1.67, depending on the domain of interest. For analyses that combine all adult respondents, this component of the unequal weighting DEFF is expected to be approximately 1.79

The third feature of the PATH Study design that contributes to the overall sampling variability is the restriction that no more than two adults be sampled from a participating household. This requirement contributes to the variability of weights because adults in some multi-person households will be sampled at lower rates than persons of the same age, race, and tobacco use

group in single- or two-person households. The unequal weighting DEFFs due to this feature of the sample design are expected to range from 1.00 to 1.02, depending on the domain of interest. For analyses that combine all adult respondents, this component of the unequal weighting DEFF is expected to be negligible (i.e., approximately equal to 1).

Note that for analyses of subgroups of race, say by age or sex, these DEFFs will diminish, because generally fewer members of the subgroups will contribute to the clustering effect.

Estimates of precision and power for the PATH Study were calculated (taking into account the DEFFs resulting from the three sample design features described previously) and are shown in Tables 3a-4b. Tables 3a and 3b are for adults and Table 4a and 4b are for youth. The projected RSEs are for a generic statistic estimating a prevalence rate of 15 percent (such as the percentage of the adult population who are every day cigarette smokers). The MDADs are for a generic statistic estimating change from a baseline prevalence rate of 10 percent (such as any non-cigarette tobacco use). Both the RSEs and MDADs presented here are for illustrative purposes.
In Tables 3a and 4a, the RSEs are for cross-sectional estimates at baseline and the MDADs are for a change from baseline to Wave 2. In Tables 3b and 4b, the RSEs are for cross-sectional estimates at Wave 4 and the MDADs are for a change from baseline to Wave 4. The subgroups of interest are defined in terms of tobacco-related behaviors, which are subject to change over time. This presents a challenge when trying to estimate the subgroup sample sizes in future waves of the PATH Study, particularly given the recent expansion of tobacco products on the market. Over time, participants sampled as youth will become young adults and those sampled as young adults (18 to 24 years of age) will move into the older age group. As a result, variation in weights among members of most subgroups will increase, and it is necessary to inflate the DEFFs assumed due to unequal weighting. It is not possible to predict the precise inflation factor for each subgroup given the complication of unknown, future rates of switching, substituting, or multiple tobacco product usage. For these reasons, one inflation factor was estimated for each follow-up wave and applied to all subgroups and the estimates of cross-sectional precision for later waves and of detectable changes across

several waves are presented for a reduced set of subgroups (i.e., for those for which the estimates are expected to be fairly robust to the assumptions made). As a consequence, the estimates herein should be interpreted with caution.

The large sample of 24,052 adult tobacco users at baseline is intended to allow analyses for many user subgroups. Table 3a highlights some subgroups of potential analytic interest by breaking out sample sizes and measures of precision and power for tobacco users, menthol smokers, users of both smoked and smokeless tobacco, and daily/non-daily tobacco users; the same statistics are shown for each of these subgroups among young adults (18 to 24 years of age) and among African Americans. Subgroup sample sizes, except those for dual users, were estimated using data from the (averaged) 2001–2007 CPS-TUSs. The definition of a "tobacco user" is therefore adopted from the CPS-TUS and is described in Section B.1b. To better reflect the substantial increase in dual use of cigarettes and smokeless tobacco in recent years, the assumed percentages of dual users amongst all users were estimated using data from the 2010 NSDUH. The RSEs for a 15 percent prevalence rate are below 8 percent for most subgroups and at or below 5 percent for more than half of them. The MDADs for a 10 percent baseline prevalence rate are almost all below 2 percentage points and for about two-thirds of the subgroups a one-year change of 1 percentage point or less can be reliably detected. For both RSEs and the MDADs, smaller is better.

Table 3a.   Baseline adult sample sizes, relative standard errors (RSEs) and minimum detectable absolute differences (MDADs) at Wave 2

Group	Baseline sample size	RSE on 15% item	MDAD on 10% item
All adults	**42,730**	**2.5**	**0.4**
Current users	24,052	2.5	0.4
Menthol smokers	6,975	3.6	0.6
Dual (smokers and smokeless tobacco users)	3,345	4.7	0.8
Daily users	19,241	2.7	0.5
Less-than-daily users	4,810	4.1	0.7
Current non-users	18,678	2.7	0.5
Adults ages 18-24	**10,709**	**3.2**	**0.8**
Current users	4,709	4.0	1.0
Menthol smokers	1,602	6.3	1.6
Dual (smokers and smokeless tobacco users)	1,157	7.3	1.9
Daily users	3,391	4.6	1.2
Less-than-daily users	1,318	6.9	1.8
Current non-users	6,000	3.6	0.9
Black/African American adults	**6,000**	**4.1**	**0.7**
Current users	2,661	5.1	0.9
Menthol smokers	1,943	5.9	1.0
Dual (smokers and smokeless tobacco users)	370	12.9	2.2
Daily users	1,969	5.9	1.0
Less-than-daily users	692	9.5	1.7
Current non-users	3,339	4.8	0.8

Table 3b suggests that the RSEs for a 15 percent prevalence rate at Wave 4 will be only slightly larger than at baseline. This is as expected because the number of households in the PATH Study was chosen to produce a youth sample of sufficient size to replenish the adult sample in future waves. A larger difference is seen when comparing the one-year versus three-year MDADs for a 10 percent baseline prevalence rate; this is due to the increased DEFFs and reduced correlations between samples by Wave 4.

Table 3b.   Adult sample sizes, relative standard errors (RSEs) and minimum detectable absolute differences (MDADs) at Wave 4

Group	Wave 4 sample size	RSE on 15% item	MDAD on 10% item
All adults	**43,064**	**2.6**	**0.6**
Current users	24,240	2.6	0.6
Menthol smokers	7,030	3.7	0.8
Dual (smokers and smokeless tobacco users)	3,372	4.9	1.1

Daily users	19,392	2.8	0.6
Less-than-daily users	4,848	4.3	0.9
Current non-users	18,824	2.9	0.6
Adults ages 18-24	**12,347**	**3.1**	**1.2**

The adult sample sizes considered in this section are based on completed interviews and therefore the estimates of precision and power apply to tobacco and health outcomes collected via the interview instrument. However, another important component of the PATH Study is the collection and analysis of biospecimens from those who consent to provide them. Blood specimens, for example, may be analyzed to detect markers of risk for tobacco-related disease. Table 2 in Section B.1f shows that 27,775 adults are expected to provide a blood sample at baseline. If one assumes that 76 percent of these adults will still be participating in the study at Wave 4 and will provide a blood sample if requested, a change in risk of about 0.42 percentage points could be detected with 80 percent power. As with all the estimates presented in this section, precision and power for finer divisions of the subgroups (e.g., by gender) are expected to be reduced.

As stated above, the initial sample of 17,070 youth at baseline is necessary both to replenish the adult cohort in future waves of the PATH Study and to provide sufficient power for analyses of youth subgroups. Table 4a shows sample sizes and measures of precision and power for the youth sample overall and by subgroups of possible interest: tobacco users, smokers, menthol smokers, "experimenters", never smokers, susceptible never smokers, and never users of tobacco; the same statistics are shown for each of these subgroups among 12 to 13 year-olds and among 14 to 17 year-olds. Subgroup sample sizes were estimated using data from the 2009 National Youth Tobacco Survey (NYTS). Experimenters were defined as youth who have ever smoked any cigarette, even one or two puffs, but fewer than 100 cigarettes. Susceptibility to initiate cigarette smoking among never smokers was defined as providing any response other than "no" to the question, "Do you think that you will try a cigarette soon?" and any response other than "definitely not" to the questions, "Do you think you will smoke a cigarette anytime during the next year?" and "If one of your best friends offered you a cigarette, would you smoke it?"

Table 4a.   Baseline youth sample sizes, relative standard errors (RSEs), and minimum detectable absolute differences (MDADs) at Wave 2

Group	Baseline sample size	RSE on 15% item	MDAD on 10% item
All youth	**17,070**	**2.7**	**0.7**
Current users	3,199	4.6	1.2
Current smokers	2,229	5.4	1.3
Menthol smokers	1,070	7.5	1.9
Experimenters	5,133	3.8	1.0
Never smokers	10,676	3.0	0.8
Susceptible never smokers	2,669	5.0	1.2
Never users	9,937	3.1	0.8
Youth ages 12 to 13	**5,609**	**3.7**	**1.4**
Current users	449	11.4	4.3
Current smokers	280	14.3	5.4
Menthol smokers	135	20.5	7.8
Experimenters	1,122	7.3	2.8
Never smokers	4,487	4.0	1.5
Susceptible never smokers	1,122	7.3	2.8
Never users	4,207	4.1	1.6
Youth ages 14 to 17	**11,461**	**3.0**	**0.9**
Current users	2,751	4.9	1.4
Current smokers	1,948	5.7	1.7
Menthol smokers	935	8.0	2.3
Experimenters	4,011	4.2	1.2
Never smokers	6,189	3.6	1.0
Susceptible never smokers	1,547	6.3	1.8
Never users	5,730	3.7	1.1

Overall, there are large samples in many of the subgroups of interest. For example, there are approximately 9,937 never users for whom tobacco use initiation rates will be tracked. Tobacco cessation is more of an issue in the older adolescent group (ages 14 to 17) and there are about 2,751 tobacco users and 1,948 cigarette smokers whose quitting behavior over time will be monitored. The smallest subgroup summarized that may be of interest is menthol smokers. If some regulatory action relating to menthol cigarettes were to be taken, these youth might respond by quitting, switching brands, or switching to other forms of tobacco use. There are 935 such participants in the 14 to 17 age range, which provides statistical power to examine all but the rarest outcomes. The RSEs for a 15 percent prevalence rate are below 8 percent for most subgroups and at or below 5 percent for more than half of them. Among all youth 12 to 17 years of age, the sample size overall and in each of the subgroups except menthol smokers is sufficient to reliably detect a one-year change of 1.5 percentage points in a 10 percent baseline

behavior overall. This is a critically important threshold because measures of quitting, initiation, and non-cigarette tobacco use tend to be in this 10 percent range (depending on the definitions used) and a statistically significant increase of one and a half percent would be of policy interest. However, Table 4b highlights the importance of recruiting a large sample of youths at baseline. For detecting change in a 10 percent baseline behavior across three years, the MDADs among all 12 to 17 year-olds are closer to 2 percentage points for several subgroups. Again, this is due to the increased DEFFs and reduced correlations between samples by Wave 4.

Table 4b.   Youth sample sizes, relative standard errors (RSEs), and minimum detectable absolute differences (MDADs) at Wave 4

Group	Wave 4 sample size	RSE on 15% item	MDAD on 10% item
All youth	**14,000**	**3.1**	**1.1**
Current users	2,624	5.4	1.9
Current smokers	1,828	6.2	2.2
Menthol smokers	877	8.7	3.0
Experimenters	4,210	4.4	1.6
Never smokers	8,756	3.5	1.2
Susceptible never smokers	2,189	5.8	2.0
Never users	8,150	3.6	1.2
Youth ages 12 to 13	**4,677**	**4.3**	**2.0**
Current users	374	13.0	6.1
Current smokers	234	16.4	7.7
Menthol smokers	112	23.6	11.0
Experimenters	935	8.4	3.9
Never smokers	3,742	4.6	2.2
Susceptible never smokers	935	8.4	3.9
Never users	3,508	4.8	2.2
Youth ages 14 to 17	**9,323**	**3.4**	**1.4**
Current users	2,237	5.7	2.4
Current smokers	1,585	6.6	2.7
Menthol smokers	761	9.3	3.8
Experimenters	3,263	4.9	2.0
Never smokers	5,034	4.2	1.7
Susceptible never smokers	1,259	7.4	3.0
Never users	4,661	4.3	1.8

B.3 Methods to Maximize Response Rates and Deal with Nonresponse

The emphasis for the baseline wave will be on maximizing the participation of selected households and selected persons in the PATH Study. For the annual follow-up waves, the focus will be on maintaining contact with respondents and maximizing their retention in the study.

Baseline Wave

The PATH Study will not be immune to the declining response rate trends experienced in recent years across most surveys. Methods to maximize response rates will be planned and implemented both prior to and during the data collection effort.

The prime contractor will recruit a team of experienced field interviewers and field supervisors sufficient in size to work all cases thoroughly. These field staff will be strategically located within or in close proximity to PSUs, which will expedite visits to the sample dwelling units and will also ensure that they are familiar with the communities within which the cases are located. Field interviewers will also be thoroughly trained in gaining respondent cooperation through refusal aversion and conversion. Field management will ensure that data collection efforts are thoroughly planned down to the field interviewer level; for example, production goals will be developed that will set a pace for individual field interviewers, field supervisor teams, and the nation as a whole.

Several tools and approaches to address nonresponse and maximize response rates will be used, in addition to the respondent incentives described in Section A.9. First, the interviews will be conducted in five languages in addition to English: Spanish, Mandarin, Vietnamese, Korean, or Cantonese; all of the instruments will be translated into these languages and bilingual field staff will administer them. Second, extensive respondent materials will be developed to encourage participation, also translated into the five non-English study languages. These materials will include an

advance letter to inform selected households of the study prior to in-person contact (Attachment 10). The advance letter will contain assurances of privacy and describe the voluntary nature of the survey and principal purposes and uses of the survey data. A PATH Study website and respondent telephone call line to answer respondents' questions and to reassure them of the credibility of the study will be established. Tailored letters will be developed for use with reluctant respondents/sample persons and with selected units located in limited-access situations (doorperson buildings, gated communities, etc.), which may be sent via FedEx or priority mail to reinforce the perceived importance of participation. (See Attachment 17 for an example of a refusal letter.) In addition, for respondents who decline to provide a blood specimen (collected with a urine specimen at a follow-up visit from a health professional), field interviewers will seek to collect the urine sample at the first home visit by offering the respondent $10 for the additional time involved.

A web-based Supervisor Management System (SMS) will allow field supervisors to closely monitor each field interviewer's work, which facilitates the development of strategies to address nonresponse. These strategies will include reassigning difficult or reluctant cases among local field interviewers and the use of specially trained, traveling field interviewers who are highly skilled in refusal conversion.

Lastly, the data collection efforts will implement a phased approach that anticipates refusal conversion efforts. In this approach, new samples of households will be released to the field interviewers approximately every 4 months. Cases from earlier 4-month data collection periods will not have to be closed out prior to releasing a new sample, which will allow additional time to complete challenging cases. Further, the most difficult-to-work segments will be released in the first or second data collection periods, thereby giving the data collection staff additional time to work the cases. Front-loading the sample release in this manner allows field interviewers the opportunity to implement the full contact strategy, including nonresponse conversion as needed.

To adjust for those non-interviews that cannot be converted, adjustments will be performed for the PATH Study data using the procedures described in Section B.2. The specific procedure selected will ensure the accuracy of the resulting estimators and the suitability of the compensated data set for addressing the major objectives of the study.

The baseline response rate (including the household screener and extended interview response rates) is expected to be 78 percent for both adults and youth. (See Section B.1 for a discussion of this expected response rate.) The response rates for the baseline extended interviews are expected to be 90 percent; these will be calculated as the number of survey respondents divided by the number of eligible sample persons. Ineligible persons include persons under the age of 12 years; respondents whose mental and/or physical impairment preclude participation in the survey; military personnel on active duty; and persons who are unable to conduct their interviews in English, Spanish, Mandarin, Vietnamese, Korean, or Cantonese.

Follow-up Waves

In the follow-up waves, the PATH Study team will seek to maintain respondent engagement as well as track respondents, so they can be contacted for follow-up data and biospecimen collection. In terms of maintaining engagement, many of the same activities conducted at the baseline wave will be completed, plus the following:

- Visit respondents who have moved up to 100 miles from a study PSU,

- Offer a web-based version of the interview for movers who are located more than 100 miles from any PSU, and

- Collect biospecimens (urine only) from movers via kits that can be mailed to the movers and returned to the biorepository.

With regard to locating respondents after the baseline wave, PATH Study staff will conduct ongoing tracking of study respondents--so they can be contacted for follow-up data and biospecimen collection--and tracing of those

who become lost at follow-up. Management of tracking/tracing activities will be supported through the home office centralized Home Management System (HMS). This component of the study management system will house the database of contact information and provide for real-time access in the field and at the home office to the most current information available. Field and home office staff involved in tracking/tracing will provide updates, and supervisors will generate reports for monitoring purposes and to determine next steps.

Using the centralized HMS as a tool, PATH Study staff will implement the following routine tracking steps to minimize the number of cases requiring intensive tracing.

- **Collect contact information at baseline for tracing references.** At baseline, respondents will be asked for the names, addresses, and telephone numbers of two people to serve as tracing references who will always know how to reach the respondent and do not live in the same household. Given that a sizeable percentage of respondents will be young adults, respondents also will be asked for information not traditionally requested (e.g., names of colleges attended) for use in social networking site searches.

- **Use interim contacts to determine if contact information has changed or if tracing is needed.** Contacts by mail or email will ask respondents to report any address changes, and the study will provide a number of easy ways this can be done, including visiting the study website, calling a toll-free number, or sending updated information via mail or email. The PATH Study also will mail to respondents stamped "address correction requested," requesting new address information for people who may have moved. In addition to supporting tracing, these interim contacts will help to maintain respondent motivation to cooperate and continued engagement with the study. PATH Study respondents will also be provided with an incentive ($5) as a thank you for updating their contact information.

- **At each in-person visit, update contact information.** During household visits for each follow-up wave, the field interviewers will update contact information on the respondent as well as on relatives or persons not living in the household who are likely to know the whereabouts of the respondent.

For those respondents lost at follow-up, PATH Study staff will implement a systematic approach to tracing their whereabouts. If the current occupants of the last known address cannot guide the field interviewer to the respondent's whereabouts, the field interviewer will carry out the first line of tracing, using the respondent's last known telephone number(s), tracing references, directory assistance, and neighbors to locate the respondent. If unsuccessful, the case will be sent to the PATH Study home office for a second line of more intensive tracing. A small team of tracers at the home office will follow protocols to trace PATH Study respondents, using tracing resources such as the following.

- **Lexis Nexis**. This database, compiled from public records, can return respondent address histories and telephone numbers. Submissions will be made at least quarterly, and the tracers will review and follow up on the results.

- **Social networking sites**. Sites such as Facebook, Google+, and Twitter offer tracing possibilities that can be quite productive, especially for persons in selected age groups. These sites will be searched, as needed.

- **Internet searches**. These searches include free and paid services. Examples of the services include online telephone directories and limited public information records.

As the need arises and the resources are available, in-person tracing (i.e., "skip tracing") will be used. This approach involves intensive in-person tracing at the respondent's last known addresses and in his/her old neighborhoods to identify contact information or current whereabouts. Because it is expensive per case, in-person tracing will be used judiciously and only after more cost-effective approaches have been attempted.

B.4 Test of Procedures or Methods to be Undertaken

 A field test will be conducted upon receipt of OMB approval. The field test will serve as a test of the data collection procedures and operations, and as a test of alternative incentive schemes and household screeners that can potentially reduce both respondent burden and study costs.

The field test will be conducted in approximately 15 PSUs that are purposively selected to reflect the diversity of PSUs selected for the main study. For example, the field test PSUs will include urban and rural PSUs, and PSUs within states that have relatively high and low tobacco use prevalence rates. The sample of households and individual respondents will be selected using the same methods planned for the main study. Potential respondents include approximately 1,823 adult (18 years and older) household screener respondents, 840 adult individual screener respondents, 590 adult extended interview respondents, 100 parent interview respondents, and 100 youth respondents (12 to 17 years). Field interviewers will obtain informed consent from field test respondents. (See Attachment 11.)

The first field test will be conducted approximately 10 months in advance of the main data collection effort, as part of the main study preparations. A field test is also planned on an annual basis in advance of the corresponding annual wave of data collection, but the scope and duration of each are expected to change year-to-year. The first field test is designed to fine-tune the data collection protocol and to inform decisions on the first-phase household screener incentive amount and length of the screener. In summary, objectives of the first field test are to test: (1) the administration and performance of the data collection instruments; (2) biospecimen collection in a household setting (buccal cells, urine, and blood specimens will be collected, packaged, shipped, and analyzed); (3) field interviewer training procedures and materials; (4) data processing and the interface between the biorepository and the prime contractor; (5) systems and security architectures; (6) alternative incentive levels for completing the household screener ($0 vs. $5 vs. $10) and incentive procedure; and (7) a short and long version of the household screener. Households will be randomly assigned to receive the three alternative incentive levels and short vs. long household screener. (See Attachment 19 for more information on the field test experiment.) The field test sample sizes have been set to provide adequate power (.60 or better) to detect effects on screener response rates from the different levels of incentive and screener lengths.

To provide for a full test of the data collection procedures and operations to be used in the main study, personally identifying and contact information will be collected from the field test respondents as it will be in the baseline data collection. As with the baseline data collection, this information will not be linked to respondent survey data or biospecimens and confidentiality procedures described in Section A.10 will be used in the field test.

Data and biospecimens collected in the field test will be analyzed to assure the quality of the data specimens collected pursuant to the instrumentation and protocol designs to meet the expectations underlying those designs, and are suitable for the analyses to meet the study objectives. OMB will be notified by way of a change request regarding any changes, substantive or non-substantive, to data collection procedures or instruments as a result of the field test and in advance of the baseline. An amendment will be submitted to OMB for each annual field test and follow-up wave of data collection to reflect any changes to the instrument or data collection procedures, and to capture corresponding estimates of burden.

B.5 Individuals Consulted on Statistical Aspects and Individuals Collecting and/or Analyzing Data

A list of individuals who consulted on statistical aspects of the PATH Study design and will collect and/or analyze the PATH Study data is included in Attachment 20.

References

Dohrmann, S., Han, D., and Mohadjer, L. (2007). Improving coverage of residential address lists in multistage area samples. Proceedings of the Joint Statistical Meetings [CD-ROM], 3219–3226. Alexandria, VA: American Statistical Association.

Iannacchione, V., Staab, J., and Redden, D. (2003). Evaluating the use of residential mailing lists in a metropolitan household survey. *Public Opinion Quarterly*, 67(2): 202–210.

Kalton, G. (1986). Handling wave nonresponse in panel surveys. *Journal of Official Statistics,* 2: 303–314.

Montaquila, J.M., Hsu, V., Brick, J.M., English, N., and O'Muircheartaigh, C. (2009). A comparative evaluation of traditional listing vs. address-based sampling frames: Matching with field investigation of discrepancies. Proceedings of the Joint Statistical Meetings [CD-ROM], 4855–4862. Alexandria, VA: American Statistical Association.

O'Muircheartaigh, C., Eckman, S., and Weiss, C. (2003). Traditional and enhanced field listing for probability sampling. Proceedings of the Joint Statistical Meetings [CD-ROM], 2563–2567. Alexandria, VA: American Statistical Association.

Rizzo, L., Kalton, G., and Brick, J.M. (1996). A comparison of some weighting adjustment methods for panel nonresponse. *Survey Methodology*, 22(1): 43–53.