Technical Documentation
2003 Detroit Arab American Study (DAAS)

Steve Heeringa, Terry Adams
Survey Research Center
Institute for Social Research
University of Michigan
April 2004

2003 DAAS Sample Design

I. STUDY POPULATION

The study population for the 2003 Detroit Arab American Study (DAAS) is defined to include all adults of Arabic or Chaldean descent who were 18 years and older and resided in households in the Detroit 3-county metropolitan area during the six-month survey period, July to December 2003. The geographic area of the survey population includes Wayne, Oakland and Macomb counties in Michigan. The survey population includes only eligible adults living in households. Individuals in institutions, living in group quarters or on military bases are excluded from the survey population.

II. DUAL-FRAME PROBABILITY SAMPLE DESIGN

The 2003 Detroit Arab American Study (DAAS) is based on a dual-frame sample design. The combined frame for the 2003 DAAS probability sample design consists of two component parts: 1) an area probability frame (Kish, 1965) used to select area segments from Census tracts in which 10% of more of persons were self-classified as of Arab- or Chaldean- American ancestry in the 2000 Census; and 2) a list frame for selecting housing units from mailing and membership lists of 13 Arab- and Chaldean- American organizations.

II.A Area Probability Sample Component

The area probability sample component of the 2003 DAAS is based on a conventional three-stage sample design, a primary stage sample of area segment units followed by a second stage sample of housing units within area segments and random selection of one eligible adult respondent in households with one or more eligible persons. The geographic domain for the area probability component of the DAAS design consisted of the 60 Census tracts in Wayne, Oakland, and Macomb counties in which at least 10% of the Census 2000 population self-identified as Arab- or Chaldean-American. These tracts included 49% of the total population of self-identified Arab- and Chaldean-Americans in the three-county area in 2000.

II.A.1 Primary Stage Sample of Area Segments. The primary stage of the 2003 DAAS area probability sample component was selected directly from computerized files extracted from the 2000 U.S. Census summary file series STF1. These files (on CD ROM) contain the 2000 Census

total population and housing unit (HU) data at the Census block level. The designated primary stage sampling units (PSUs), termed "area segments", are comprised of single Census blocks or combinations of Census blocks. Each area segment was assigned a measure of size (MOS) equal to the total 2000 occupied housing unit count for its geographic area. In creating the DAAS primary stage area segments, Census blocks were linked prior to sample selection to create area segments that included a minimum count of 75 occupied housing units.

From the geographic domain of 60 tracts, 80 primary stage area segments were sampled with probability proportionate to size (PPS).

To reduce costs of sample development and avoid placing field staff in the sample neighborhoods before the official launch of the project data collection, a decision was made to avoid the costs of a traditional enumerative housing unit listing for the selected 80 segments. Instead, a data set of United States Postal Service deliverable addresses for the ZIP code areas that incorporated all 60 Census tracts in the are probability domain of the DAAS sample was ordered from the MelissaData commercial marketing service. This large database of USPS address listings was then submitted to Geographic Data Technologies for geocoding of Census tracts and blocks. The resulting geocoded database was then merged against the selected Census tract and block identifiers for the 80 selected area segments to produce a housing unit listing for each area segment.

The listings for one-half of the 80 area segments were randomly selected for review by SRC field staff that visited the segments prior to the start of data collection and checked to ensure that the postal address listings were complete and accurate. If errors in the USPS housing unit listings or associated geocoding for these updated segment listings were identified, they were corrected. Listings for the remaining 40 area segments in the primary stage sample were not updated prior to the survey but incorrect addresses were identified in the screening process and coded as nonsample cases.

II.A.2 Second Stage Selection of Housing Units from the Area Segment Listings

A second stage sample of housing units was then selected from each primary stage area segment. The second stage sampling rates for selecting households in the DAAS area probability sample segments were computed using the following “selection equation”:

$$f = f_1 \times f_2$$

$$= \frac{MOS_{segment} \cdot a}{MOS_{domain}} \times \frac{C}{MOS_{segment}}$$

where: f = the overall multi-stage sampling rate for housing units,
=.039891 for the DAAS area probability sample;
$MOS_{segment}$ = total occupied HUs in the area segment;

$MOS_{domain,}$ = total occupied HUs in the area sample geographic domain;

a = number of area segments selected = 80;

C = average expected cluster size per segment= ($f \times MOS_{domain}/a$).

The second stage sampling rate for selecting an equal probability sample from the listed housing units for the area segment was therefore:

$$f_2 = \frac{f}{f_1} = \frac{f \times MOS_{domain}}{a \times MOS_{segment}}$$

The second stage sampling rate was computed for each selected area segment in the DAAS area probability sample design. This rate was then used to select a systematic random sample of actual housing units from the area segment listing. Selected housing units were designated for contact and screening to determine if a household member would be eligible for the study interview.

The 2003 DAAS area probability sample component yielded an equal probability sample of n=3352 listed housing units. The overall probability of selection for DAAS area probability sample households was f=0.039891 or 39.891 in 1000.

II.B The 2003 DAAS List frame Sample Component

The DAAS list frame was constructed from member or participant listings received from organizations including community, religious, business, educational and social organizations that serve the Detroit Arab- or Chaldean-American populations. Access to the member lists was facilitated by the members of these populations who served on the DAAS advisory committee. The size of the lists ranged from 350 to 10,500 names. Of the 13 lists finally used, eight were in electronic format. Five lists were received in paper printout format; of these, four were small enough to be key-entered and one (with 8000 names) was scanned into electronic format and then edited. The consolidated file included 33,841 entries.

After initial cleaning of the consolidated file to remove listings with no apparently deliverable mail address, a revised subset of 33,417 usable addresses, with names, was submitted to Lorton Data for National Change of Address (NCOA) processing. NCOA processing provides both a standardized version of the original address (with standard spellings, corrected ZIP codes, and carrier route and other postal delivery fields), and new addresses where a Change of Address form was filed by the family or individual in the previous three years. After additional cleaning and change of address steps, the final list frame for the 2003 DAAS included 29,879 apparently deliverable residential addresses in Wayne, Oakland and Macomb counties. These addresses were submitted to Geographic Data Technologies for geocoding of Census tract and block identifiers. The geocoded database was then matched against the list of Census tracts that comprised the DAAS

area probability sample domain. Addresses matched to a Census tract in the area probability domain were removed from the master data set, leaving a total of 10,645 address listings from the balance of the geographic area in the Wayne, Oakland and Macomb county survey population.

Prior to the sample selection, the housing units addresses in the frame were sorted geographically. A systematic random sample of individual addresses was then selected from the ordered list. The result was an equal probability sample of 1775 address listings selected from these 10,645 frame listings for an overall sampling fraction of f=.166745.

For purposes of interviewer assignments and efficiency, the 1775 list-frame addresses were grouped into geographic clusters of proximal addresses. These geographic clusters were termed administrative clusters and used in much the same fashion as area segments for purposes of interviewer work assignments.

II.C Third stage sampling of eligible respondents:

Each sample housing unit in the dual frame DAAS sample (area probability and list) was contacted in person by a member of the Survey Research Center's trained interviewing staff. No substitution of sample addresses or other non-probability sample selection methods were permitted. Within each cooperating sample housing unit, the SRC interviewer completed the sample cover sheet (see Appendix A) and conducted a short screening interview with a knowledgeable adult to determine if household members met the study eligibility criteria. If the informant reported that one or more eligible adults lived at the sample housing unit address, the interviewer prepared a complete listing of household members and proceeded to select a random respondent for the study interview. The random selection of the respondent was performed using a special adaptation of the objective household roster/selection table method developed by Kish (1949).

III. 2003 DAAS SAMPLE DESIGN ASSUMPTIONS, SPECIFICATIONS AND OUTCOMES

The 2003 DAAS was designed to target a total sample of 1000 completed interviews, 500 interviews with respondents in the area probability sample domain and 500 interviews with eligible respondents from other Census tracts in Wayne, Oakland and Macomb counties. Table 1 below compares the original (pre-survey) sample design specifications and assumptions to the actual 2003 DAAS outcomes for the area probability sample, list frame sample and combined 2003 DAAS dual-frame sample.

A total sample of 5127 sample housing units addresses was selected for the 2003 DAAS, 3352 listings from the area probability frame and 1775 address from the list sample frame. A total of 4619 households were contacted for screening. The actual occupancy rate for the area probability sample was 0.91 compared to a design expectation of 0.90. The

occupancy/contactable address rate for the 1775 lines in the list sample was .88, which was lower than the design expectation of 0.94. This discrepancy can be attributed to inexperience with this particular list frame and overestimation in setting the original design expectation.

Screening response rates for the 2003 DAAS were high, exceeding design-stage expectations in both the area probability and list sample components. Screening interviews were completed with 96% of area sample households and 97% of households contacted in the list sample. Among screened households, the prevalence of eligible Arab- or Chaldean-American households was lower than the design-stage estimates for both the list (actual, 52.3% vs. expected, 76.4%) and the area probability sample (actual, 20.4% vs. expected, 25.6%). The significant difference in the list sample eligibility can again be attributed to inexperience with the particular list and poor design-stage estimation of the true eligibility rate. Interviewers who worked the list frame cases often encountered new occupants at the sample address and were voluntarily told that a household of Mideastern ancestry had been the previous owner/occupant. However, since for confidentiality reasons the list frame was treated as a sample of housing units and not a sample of named individuals or families, the sample design protocol did not permit interviewers to track previous occupants to their new address.

A total of 1389 eligible households were identified in the 2003 DAAS sample household screening. Of these 1016 (73.1%) completed the study interview-- 446 interviews were completed with eligible households selected from the area probability sample frame and 570 interviews were completed with eligible list frame households.

The final response rates for the 2003 DAAS were computed based on the American Association for Public Opinion Research (AAPOR) standard: http://www.aapor.org . The final row of Table 1 presents the final AAPOR RR3 response rate calculation for the 2003 DAAS area probability (73.8%), list frame (73.3%) and combined dual–frame samples (73.7%).

Table 1: Sample Design Specifications and Assumptions. 2003 Detroit Arab-American Study.

Item	2003 DAAS Dual-Frame Sample		Area Probability Component		List Frame Component	
	Expected	**Actual**	**Expected**	**Actual**	**Expected**	**Actual**
Completed Interviews	1000	1016	500	446	500	570
Interview Response	0.74	0.73	0.72	0.75	0.75	0.72
Eligible Sample Households	1360	1389	694	595	666	794
Eligibility Rate	0.380	0.313	.256	.204	.764	.522
Screened Households	3583	4438	2712	2919	871	1519
Screening Response	0.90	.96	0.90	0.96	0.90	0.97
Occupied Households	3982	4619	3014	3048	968	1571
Occupancy Rate	.91	.90	0.90	0.91	0.94	0.88
Total Sample Housing Units	4372	5127	3348	3352	1024	1775
AAPOR RR3 Response Rate	-	73.1%	-	72.8%	-	73.3%

IV. WEIGHTED ANALYSIS OF 2003 DAAS DATA

The 2003 DAAS data set includes person-level analysis weights that incorporate sample selection, nonresponse and post-stratification factors. This weight should be used in computing estimates of descriptive statistics for the survey population (e.g. estimates of population means and proportions) and is also recommended for estimation of analytical statistics (e.g. regression coefficients, odds ratios) required in modeling relationships among variables in the survey population.

IV.A CONSTRUCTION OF ANALYSIS WEIGHTS

IV.A.1 Sample Selection Weight

The dual-frame probability sample design for the 2003 DAAS results in an unequal probability sample of households. Sample households in the area probability frame were selected with a probability of f_{AP}=.039891 while list frame sample households were chosen with probability f_{LIST}=.166745. Within sample households a single adult respondent was chosen at random to be interviewed. Since the number of eligible adults may vary from one household to another, the random selection of a single adult introduces inequality into respondents' selection probabilities. In analysis, a respondent selection weight should be used to compensate for these unequal selection probabilities. The value of the respondent selection weight is equal to the reciprocal of the household selection probability multiplied by the number of eligible adults in the household from which the random respondent was selected.

For area probability sample respondents, the selection weight factor is:

$$W_{sel} = (1/.039891) \times (\#\, Eligible \;\; Adults).$$

For list frame sample cases, the selection weight factor is:

$$W_{sel} = (1/.166745) \times (\#\, Eligible \;\; Adults)$$

In theory, the sample selection weight factor could also include a small adjustment for new housing units identified in the detailed update half-sample of 40 of a total of 80 area probability segments. Due to budget and time limitations required to code and compile the updated entries, this address listing update factor is not included in the sample selection weight.

IV.A.2 Household Nonresponse Adjustment Factor

The most widely accepted approach to compensating for nonresponse after the survey data has been collected is to develop and apply a nonresponse adjustment factor to the weight variable that is used in analysis. Underlying weighting for unit nonresponse is a model of the response propensity –the probability that the unit will cooperate in the survey request. In a sense, the concept of response propensity treats response to the survey as another step in the "sample

selection process". But unlike true sample selection in which the sampling statistician pre-determines the sampling probability for each unit, an underlying propensity model-- for the most part outside the control of a statistician-- determines the probability that a sampled case will be observed. The multiplication of the original sample selection weight for each sample unit by the reciprocal of its modeled response propensity creates a new weight, which if the model is correct enables unbiased estimation of population statistics from the survey data.

Modeling response propensity requires observations of the predictor (or independent) variables in the model for both respondents and nonrespondents. In new cross-sectional sample surveys such as the DAAS, this limits the nonresponse adjustment to characteristics of respondents and households that are known from the sampling frame or are completely observed in the screening process. Since demographic variables such as age and gender were not obtained for many DAAS nonrespondents, the nonresponse adjustment model was developed using geographic data that were available for all cases on the sample frame. The nonresponse adjustment procedure used for the DAAS is labeled the "weighting class method" (Little and Rubin, 2002). Under this method, area probability and list sample cases were grouped according to the area segment or administrative cluster (list sample) definitions. If one of these geographic "cells" contained fewer than 15 cases, it was combined with a neighboring geographic cell. DAAS area probability sample cases were assigned to thirty six (36) geographic cells; list sample cases were grouped by administrative clusters to form 20 weighting class adjustment cells.

The weighting class method makes the simple assumption that the response propensities for cases within a given weighting class cell are equal (MAR-missing at random). The common response propensity for cases in a cell is estimated by the empirical response rate for the cases assigned to that cell. DAAS control file data were used to compute the response rate for each of the 56 geographic weighting class cells. The nonresponse adjustment factor for each sample case's analysis weight is the reciprocal of the response rate for its assigned cell:

$$W_{nr,i} = 1/RRate_{cell \subset i}$$

where: $RR_{cell \subset i}$ = the response rate for the cell to which the i^{th} case is assigned.

IV.A.3 Post-Stratification Factor

The nonresponse adjustment procedures described above have the property that only data for sampled respondent and nonrespondent cases were used to compute weighting adjustments.. Another weighting technique to improve the quality of sample survey estimates is to bring in known information on the full population—both sampled units and those that were not sampled. Post-stratification is one such method for using population data in survey estimation. Simple post-stratification involves adjusting the final weights for sample cases so that weighted sample distributions conform to known distributions for the population that the sample is designed to represent.

The post-stratification factor applied to each respondent weight is computed as:

$$W_{ps,l,i} = \frac{N_l}{\sum_{i=1}^{n_l} (W_{sel,i} * W_{nr,i})}$$

where: N_l = a known population count for post-stratum l=1,…,L.

In many surveys, detailed demographic data from the 2000 Census of Population or large sample estimates of population characteristics based on the Current Population Survey (CPS) are used to develop detailed post-stratification factors for the analysis weight. Given the uncertainty over detailed Census demographic counts for the Arab- and Chaldean-American populations of interest in the DAAS, a detailed post-stratification of the final analysis weights was not applied. Instead, a very simple post-stratification weight was developed that only adjusted the weights to 2000 Census population counts for the two major domains of the survey population: the populations living in Census tracts with 10% + Arab or Chaldean population; and the balance of the survey population that lives in all other Census tracts in Wayne, Oakland and Macomb counties.

The post-stratification factor applied to the interviewed cases in the area probability sample is: $W_{ps,l,i} = 498/36241$. The post-stratification factor applied to each interviewed case from the list sample frame is: $W_{ps,l,i} = 518/9777$. These factors scale the nonresponse-adjusted sample selection weights so that the weighted total for interview cases in the area probability domain is 498 (49%) and the weighted total for cases from the balance of the metro area is 518 (51%), the percentages of the combined weighted sample total matching the 2000 Census percentages for these two domains.

IV.B FINAL ANALYSIS WEIGHTS

The final analysis weight for each 2003 DAAS respondent is computed as the product of the three weight components: $W_{final,i} = W_{sel,i} \times W_{nr,i} \times W_{ps,l,i}$;

where: $W_{sel,i}$ = the selection weight factor for respondent i=1,…,n;
$W_{nr,i}$ = the nonresponse weight adjustment factor for respondent i=1,…,n; and
$W_{ps,l,i}$= the post-stratification factor for respondent i=1,…n.

The final analysis weights are the product of the selection weight, the nonresponse adjustment factor and the post-stratification factor. The final analysis weight for the 2003 DAAS is found in the variable FINALWGT. The weight is "centered" so its sum across interview cases is equal to 1016—the number of DAAS interviews. The analysis weight values range from a minimum value of 0.370724 to a maximum value of 2.755778. The mean of this

"centered" weight is 1.0, the median is 0.909613, and coefficient of variation of the weight values is CV(FINALWGT)=.4793.

V. PROCEDURES FOR SAMPLING ERROR ESTIMATION

The 2003 DAAS was based on a dual-frame probability sample of Detroit metropolitan area households. Although smaller in scale, the DAAS sample design is very similar in its basic structure to the multi-stage designs used for major federal survey programs such as the Health Interview Survey (HIS) or the Current Population Survey (CPS). The survey literature refers to the DAAS, HIS and CPS samples as complex designs, a loosely-used term meant to denote the fact that the sample incorporates special design features such as stratification, clustering and differential selection probabilities (i.e., weighting) that analysts must consider in computing sampling errors for sample estimates of descriptive statistics and model parameters. This section of the 2003 DAAS sample design description focuses on sampling error estimation and construction of confidence intervals for survey estimates of descriptive statistics such as means, proportions, ratios, and coefficients for linear and logistic regression models.
Standard programs in statistical analysis software systems such SAS, SPSS and STATA assume simple random sampling (SRS) or equivalently independence of observations in computing standard errors for sample estimates. In general, the SRS assumption results in underestimation of variances of survey estimates of descriptive statistics and model parameters. Confidence intervals based on computed variances that assume independence of observations will be biased (generally too narrow) and design-based inferences will be affected accordingly.

V.A Sampling Error Computation Methods and Programs

Over the past 50 years, advances in survey sampling theory have guided the development of a number of methods for correctly estimating variances from complex sample data sets. A number of sampling error programs which implement these complex sample variance estimation methods are available to DAAS data analysts. The two most common approaches to the estimation of sampling error for complex sample data are through the use of a Taylor Series linearization of the estimator (and corresponding approximation to its variance) or through the use of resampling variance estimation procedures such as Balanced Repeated Replication (BRR) or Jackknife Repeated Replication (JRR). New Bootstrap methods for variance estimation can also be included among the resampling approaches. See Rao and Wu (1988).

V.A.1 Taylor series linearization method:

When survey data are collected using a complex sample design with unequal size clusters, most statistics of interest will not be simple linear functions of the observed data. The linearization approach applies Taylor's method to derive an approximate form of the estimator that is linear in statistics for which variances and covariances can be directly and easily estimated (Woodruff, 1971). SUDAAN, STATA and now SAS V8.2/V9.0 are commercially available statistical software packages that include procedures that apply the Taylor series

method to estimation and inference for complex sample data.

SUDAAN (Shah et al., 1996) is a commercially available software system developed and marketed by the Research Triangle Institute of Research Triangle Park, North Carolina (USA). SUDAAN was developed as a stand-alone software system with capabilities for the more important methods for descriptive and multivariate analysis of survey data, including: estimation and inference for means, proportions and rates (PROC DESCRIPT and PROC RATIO); contingency table analysis (PROC CROSSTAB); linear regression (PROC REGRESS); logistic regression (PROC LOGISTIC); log-linear models (PROC CATAN); and survival analysis (PROC SURVIVAL). SUDAAN V7.0 and earlier versions were designed to read directly from ASCII and SAS system data sets. The latest versions of SUDAAN permit procedures to be called directly from the SAS system. Information on SUDAAN is available at the following web site address: http://www.rti.org. Programs in SAS Version 8.2 and higher (PROC Surveymeans, PROC SurveyReg) also use the Taylor series method to estimate variances of means and regression model coefficients from complex sample survey data. New programs that will be included in a future release of SAS version 9 will permit survey-based analysis of contingency tables and logistic regression models.

An example SUDAAN command setup for estimating variances of statistics estimated from the DAAS data set is as follows:

```
PROC {procedure name}  FILETYPE=SAS DESIGN=WR;
NEST {stratum variable name} {cluster variable name};
WEIGHT {weight variable name};
….
….
RUN;
```

Stata (StataCorp, 1997) is a more recent commercial entry to the available software for analysis of complex sample survey data and has a growing body of research users. STATA includes special versions of its standard analysis routines that are designed for the analysis of complex sample survey data. Special survey analysis programs are available for descriptive estimation of means (SVYMEAN), ratios (SVYRATIO), proportions (SVYTOT) and population totals (SVYTOTAL). STATA programs for multivariate analysis of survey data include linear regression (SVYREG), logistic regression (SVYLOGIT) and probit regression (SVYPROBT). STATA program offerings for survey data analysts are constantly being expanded. Information on the STATA analysis software system can be found on the Web at: http://www.stata.com.

V.A.2 Resampling methods:

BRR, JRR and the bootstrap comprise a second class of nonparametric methods for conducting estimation and inference from complex sample data. As suggested by the generic label for this class of methods, BRR, JRR and the bootstrap utilize replicated subsampling of the

sample database to develop sampling variance estimates for linear and nonlinear statistics. WesVar PC (Brick et al., 1996) is a software system for personal computers that employs replicated variance estimation methods to conduct the more common types of statistical analysis of complex sample survey data. WesVar PC was developed by Westat, Inc. and is distributed along with documentation free of charge to researchers from Westat's Web site: http://www.westat.com/wesvarpc/. WesVar PC includes a Windows-based application generator that enables the analyst to select the form of data input (SAS data file, SPSS for Windows data base, dBASE file, ASCII data set) and the computation method (BRR or JRR methods). Analysis programs contained in WesVar PC provide the capability for basic descriptive (means, proportions, totals, cross tabulations) and regression (linear, logistic) analysis of complex sample survey data. WesVar also provides the best facility for estimating quantiles of continuous variables (e.g. 95%-tile of diastolic blood pressure) from survey data. WesVar Complex Samples 4.0 is the latest version of WesVar PC. Researchers who wish to analyze the 2003 DAAS data using WesVar PC should choose the BRR or JRR (JK2) replication option.

Another software option for the estimation of sampling errors for survey statistics in the IVEWare system. IVEWare has been developed by the Survey Methodology Program of the Survey Research Center and is available free of charge to user at: http://www.isr.umich.edu/src/smp/ive/ . IVEWare is based on SAS Macros and requires SAS Version 6.12 or higher. The system includes programs for multiple imputation of item missing data as well as programs for variance estimation in descriptive (means, proportions) and multivariate (regression, logistic regression, survival analysis) analysis of complex sample survey data.

These new and updated software packages include an expanded set of user friendly, well-documented analysis procedures. Difficulties with sample design specification, data preparation, and data input in the earlier generations of survey analysis software created a barrier to use by analysts who were not survey design specialists. The new software enables the user to input data and output results in a variety of common formats, and the latest versions accommodate direct input of data files from the major analysis software systems.

V.B Sampling Error Computation Models

Regardless of whether the linearization method or a resampling approach is used, estimation of variances for complex sample survey estimates requires the specification of a *sampling error computation model*. DAAS data analysts who are interested in performing sampling error computations should be aware that the estimation programs identified in the preceding section assume a specific sampling error computation model and will require special sampling error codes. Individual records in the analysis data set must be assigned sampling error codes that identify to the programs the complex structure of the sample (stratification, clustering) and are compatible with the computation algorithms of the various programs. To facilitate the computation of sampling error for statistics based on 2003 DAAS data, design-specific sampling error codes will be routinely included in all public-use versions of the data set. Although minor recoding may be required to conform to the input requirements of the individual programs, the sampling error codes that are provided should enable analysts to conduct either Taylor Series or

Replicated estimation of sampling errors for survey statistics.

Two sampling error code variables are defined for each case based on the sample design primary stage unit (PSU) and area segment or administrative cluster in which the sample household is located.

Sampling Error Cluster Code (CLUSTER) and Stratum Code (STRATUM). In variance estimation for complex sample designs, the sampling error clusters represent the "ultimate clusters" (Kalton, 1977) of the sample selection process. The CLUSTER code reflects the geographic clustering of sample observations based on the area segments or administrative clusters to which they are assigned. Sampling error calculation clusters for the 2003 DAAS were formed by first ordering the area segments and administrative clusters of the dual frame sample design by sample component (area probability, list frame) and within sample component by the original geographic stratification used in the sample selection. Following this ordering, area segments or administrative clusters were assigned to explicit sampling error calculation strata (STRATUM). The geographically ordered assignments of the sampling error calculation strata ensured that each stratum had a minimum of approximately 24 sample observations. A total of 32 sampling error calculation strata were formed.

Within each sampling error calculation stratum, area segments and administrative clusters were randomly assigned to one of two sampling error calculation clusters (CLUSTER). Variances are therefore estimated under the assumption that two combined PSU sampling error clusters were selected from each stratum. The combining of area segments and list-frame administrative clusters to form sampling error clusters is a necessary step to protect against possible geographic identification and disclosure for individual sample clusters. Equally important, combining area segments to form sampling error clusters ensures a minimum number of DAAS observations per cluster (here a minimum of 12 to 15). This minimum size criteria for the combined sampling error calculation clusters protects against the occurrence of "sampling zeros" in clusters for those analyses in which the researcher is focusing only on a subpopulation (e.g. women age 50 and older) of the respondents. Combining observations to form sampling error clusters in this fashion is a standard practice. Estimates of variance computed under this sampling error calculation model remain unbiased (Kalton, 1977).

References

Binder, D.A. (1983), "On the variances of asymptotically normal estimators from complex surveys," *International Statistical Review,* Vol. 51, pp. 279-292.

Brick, J.M., Broene, P., James, P., & Severynse, J. (1996). "A User's Guide to WesVar PC." Rockville, MD: Westat, Inc.

Cochran, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons.

Cohen, S.B. (1997). "An evaluation of alternative PC-based software packages developed for the analysis of complex survey data," *The American Statistician*, Vol. 51, No. 3, pp. 285-292.

Kalton, G. (1977), "Practical methods for estimating survey sampling errors," *Bulletin of the International Statistical Institute*, Vol 47, 3, pp. 495-514.

Kish, L. (1949). "A procedure for objective respondent selection within the household," .

Kish, L. (1965), *Survey Sampling*. New York: John Wiley & Sons, Inc.

Kish, L., & Hess, I. (1959), "On variances of ratios and their differences in multi-stage samples," *Journal of the American Statistical Association*, 54, pp. 416-446.

LePage, R., & Billard, L. (1992), *Exploring the Limits of Bootstrap*. New York: John Wiley & Sons, Inc.

Rao, J.N.K & Wu, C.F.J. (1988.), "Resampling inference with complex sample data," *Journal of the American Statistical Association*, 83, pp. 231-239.

Research Triangle Institute (2001). SUDAAN User's Manual, Release 8.0. Research Triangle Park, NC: Research Triangle Institute.

Rust, K. (1985). "Variance estimation for complex estimators in sample surveys," *Journal of Official Statistics*, Vol. 1, No. 4.

SAS Institute, Inc. (2003). SAS/STAT® User's Guide, Version 9, Cary, NC: SAS Institute, Inc.

Shah, B.V., Barnwell, B.G., Biegler, G.S. (1996). SUDAAN User's Manual: Software for Statistical Analysis of Correlated Data. Research Triangle Park, NC: Research Triangle Institute.

Skinner, C.J., Holt, D., & Smith, T.M.F. (1989). *Analysis of Complex Surveys*. New York: John Wiley & Sons.

SPSS, Inc. (1993). SPSS® for Windows™: BASE System User's Guide, Release 6.0. Chicago, Il: SPSS Inc.

STATA Corp. (2001). STATA Statistical Software: Release 7.0. College Station, TX: STATA Corporation.

Westat, Inc. (2000). WesVar 4.0 User's Guide. Rockville, MD: Westat, Inc.

Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Woodruff, R.S. (1971), "A simple method for approximating the variance of a complicated estimate," *Journal of the American Statistical Association*, Vol. 66, pp. 411-414.