Methodology for the Princeton RDD Survey
August, 2008

Background

In the summer of 2008, Westat conducted a national random digit dial (RDD) survey on behalf of Princeton University. Major topics covered in the survey included worker perceptions of the off-shoreability of their jobs, occupational licensing, and adult lifetime work experience. This document discusses the methods used by Westat to conduct the survey.

Questionnaire Development

Princeton provided Westat with a draft of the desired questionnaire at the start of the project. Princeton and Westat staff collaborated in finalizing the question order and wording. A number of the questions had been developed and tested in earlier work by Princeton and under prior task order contracts with Westat. Several questions regarding the respondent's employer, job activities, and demographics, were taken from the Current Population Survey. Westat programmed the questionnaire and skip patterns for administration by Computer Assisted Telephone Interviewing (CATI), in both English and Spanish. A few days before data collection, Westat staff pretested the instrument with several volunteer respondents. This pretest suggested several additional revisions for the questionnaire, including shortening it so as to achieve the desired average interview length of 15 minutes.

The CATI interview began with a short screener instrument designed by Westat in order to determine the eligibility of household adults for the extended interview (i.e., the study questionnaire). In order to be eligible for the study, persons had to be adults in the labor force – for this project the labor force was defined as persons who are either a) currently working at a job for pay or profit, or b) currently looking for work and have worked at a job in the past. Households in which no adults are currently in the labor force were not eligible for the study. If a household contained more than one adult in the labor force, one was randomly selected by the CATI program for participation in the extended interview. When the person chosen for the extended interview was someone other than the screener respondent, age and work force status was confirmed with these persons before continuing with the extended interview.

Sampling and Data Collection

Standard "list-assisted" random digit dialing (RDD) techniques were used to identify and select respondents for the survey. Under the list-assisted RDD approach, a sampling frame of 100-number banks is created from all known area codes and telephone exchanges in the United States, where a 100-number bank is defined by the three-digit

area code, three-digit exchange, and next two digits of the telephone number. The frame is then matched against published residential telephone directories, and a count of the number of listed residential numbers was determined for each 100-number bank.. As is the practice in most list-assisted RDD surveys conducted by Westat, the working banks (i..e., those remaining in the sampling frame) were those with 1 or more listed residential phone numbers. While we could have gained operational efficiency by limiting the working banks to those with a greater number of listed residential phone numbers, doing so would increase potential undercoverage since households in the excluded banks would have no chance of being selected for the survey.

An equal probability sample of telephone numbers was then drawn from the frame of working banks described above. Relatively inexpensive automated procedures were initially be applied to the selected numbers to remove known nonresidential and nonworking numbers, leaving for screening and subsequent processing those numbers that are more likely to be active residential numbers. An initial sample of 71,000 telephone numbers was drawn, along with a reserve sample of an extra 20,000 numbers. During data collection the reserve sample was released to be worked, as it became clear that this was necessary to achieve the desired target of 2,500 completed interviews.

Data collection lasted approximately six weeks, beginning on June 5th, 2008 and ending on July 20th. Westat trained approximately 100 interviewers for the project, but as data collection proceeded less productive interviewers were dropped from the study. We completed 2,513 interviews. Table 1 presents details on the outcomes of our data collection and the response rates.

Table 1. Outcomes of Data Collection

	Estimated rate	Number in sample
Telephone numbers (list-assisted method)	——	91,000
Numbers determined to be nonresidential via business purge	3.7%	3,409
Numbers available for telephone screening		87,591
Finalized cases for which residential status is undetermined	15.3%	13,437
Ring no answer, no contact	74.1%	9,958
Answering machine	25.8%	3,462
Other nonresponse	0.1%	17
Finalized cases for which residential status is determined	84.7%	74,154
Determined to be nonresidential	75.0%	55,634
Determined to be residential (households)	25.0%	18,520
Households completing screener	33.3%	6,165
Ineligible households	11.3%	2,086
Nonresponding households	66.7%	12,355
Households available for extended interviewing		4,079
Completed extended interviews	61.6%	2,513
Eligible nonrespondents to extended interview	26.2%	1,067
Ineligible	1.9%	76
Unknown eligibility	10.4%	423
Screener Response Rate	28.2%	
Conditional Extended Interview Response Rate	63.5%	
Final Overall Response Rate	17.9%	

Weighting the Survey Data

The RDD survey was weighted to compensate for variable selection probabilities, differential response rates, and possible undercoverage of the sampling frame. The weighting process was carried out in five steps as described below.

Base weights for sample telephone numbers. Since the sampling process started with the selection of telephone numbers, a base weight was first attached to each selected telephone number, followed by weighting of subsequent sampling units (i.e., households

and persons). The base weight for a sampled telephone number is the inverse of the selection probability of the telephone number .

Residential status adjustment. Only residential telephone numbers are of interest and they were identified through screening. Since it was not always possible to determine residential status, the weight of unknown cases (telephone numbers with unknown residential status) was distributed so as to preserve the distribution of the cases for which residential status was known. This adjustment was carried out as a nonresponse adjustment within a number of adjustment cells defined by the Census Region, metropolitan statistical area (MSA) status, and minority status. All of these variables are exchange-level variables available in the RDD sampling frame. A cross-classification of these variables yielded a total of 16 adjustment cells.

- 4 Census Regions: (1) Northeast, (2) Midwest, (3) South, and (4) West;

- 2 MSA statuses: (1) MSA, and (2) nonMSA;

- 2 minority-status categories: (1) low minority and (2) high minority, where the low minority group is defined as those telephone numbers belonging to the exchanges with less than 60 percent white population.

Eligibility status adjustment. There were 18,520 residential telephone numbers screened to be residential. Among them, 4,079 households had eligible persons, and 2,086 did not have an eligible person - these were households that had no adults in the labor force at the time of interview. As anticipated, for some residential telephone numbers (12,355), it was not possible to ascertain eligibility status. Therefore, an eligibility status adjustment was performed using new adjustment cells defined by the Census Region, MSA status, and median income of the telephone exchange. Five median income categories were defined, and there were altogether 50 adjustment cells.

The household screening questionnaire determined how many eligible adults lived in the household. There are three categories of interest. The first category includes households in which one or more residents are eligible for the extended interview. The second category is households in which no resident is eligible. The third category includes households in which eligibility is not known. It is possible that some of the unknown cases were eligible. Therefore, the weight of the eligible cases was adjusted upwards within each of the adjustment cells defined above,

Adjustment for multiple telephone households and sample person base weight. A household with multiple telephones had a higher probability to be selected in proportion to the number of telephones, so this should be reflected in weighting. Let k_i be the number of telephone numbers associated with household *i*. The eligibility adjusted weight obtained in step 3 was then adjusted by dividing it by k_i. However, k_i was capped at 3 (i.e., $k_i = 3$ for households with 4 or more telephones in use) to avoid too much variability in weights.

Only one person was selected for an extended interview from all eligible persons in a participating household. If m_i is the number of eligible persons in household *i*, then the probability of selecting an eligible person from the household is $1/m_i$. The weight attached to the sample person from household *i* is then given by:

$$w_i^{(pre)} = w_i^{(re)} \frac{m_i}{k_i}$$

The value of m_i was capped at 4 to avoid excessive variability in weights. There were 4,079 households involved in this step of weighting.

Person-level nonresponse adjustment. This adjustment compensated for nonresponse resulting from sample persons who agreed to participate in the study but for some reason did not complete the extended interview. Furthermore, some people were found ineligible during the extended interview, and also there were some nonrespondents whose eligibility could not be confirmed. To reflect these response statuses, a similar adjustment done in the first step described above was performed using the weighting cell approach. The weighting cells were first defined by the following variables:

- Census Region (1, 2, 3, 4);
- MSA status (MSA, non-MSA);
- Median income level of exchange (5 categories);
- Employment status (employed, unemployed).

Note that the first three variables are exchange-level variables, whereas the last variable is a person-level variable available from the screening interview. Six final cells were determined by a CHAID analysis as shown in table 2. CHAID stands for Chi-squared Automatic Interaction Detector, which performs cell collapsing according to user criteria. We specified the minimum cell size to be 30 and cells were merged if the difference in cell response rates was not significant at the 5 percent significance level, based on the Chi-squared test. Note that MSA status did not enter in the formation of the final cells.

Table 2. The Final Adjustment Cells Determined by CHAID

Cell Number	Region	Median Income	Employment Status	Counts
1	1 and 3	all	Employed	1,972
2	2 and 4	all	Employed	1,776
3	all	1 and 2	Unemployed	146
4	all	3	Unemployed	73
5	all	4	Unemployed	68
6	all	5	Unemployed	44
Total				4,079

Calibration Adjustments. Finally, the nonresponse-adjusted person weights were calibrated to known population control totals obtained from the June 2008 Current Population Survey (CPS) data. This was done using a raking algorithm. The marginal dimensions of the raking procedure were defined by the following variables (CPS source variable names are given in parentheses).

- Sex (PESEX);
- Age group (PRTAGE): 18-35, 36-49, 50+;
- Educational attainment (PREDUCA4): Some high school, some college;
- Census Region (GEREG);
- MSA status (GTMETSTA): MSA, non-MSA;
- Race/ethnicity (PTDTRACE and PEHSPNON): recoded into two groups, non-Hispanic white only and all others;
- Employment status (PRFTLF and PREMPNOT): full time employed/part time employed/unemployed;
- Type of Employer (PRCOWPG): private / non-private.

Some of these variables were combined, and altogether 5 raking dimensions were defined as shown in table 3.

Table 3. Five Raking Dimensions

Raking Dimension	Number of Categories
Sex by Age group	6
Education level	2
MSA by Region	8
Race/ethnicity by Employment status	6
Employment type	2

We defined the corresponding variables using the survey data. However, some of the variables involved were survey questions that had resulted in some degree of missing values. Before running the raking algorithm, the missing values were imputed using hot-deck estimation procedures, which use similar values from the complete records in the data. The resulting weights of the raking procedure are the final weights to be used in analysis.

Industry, Occupation, and Off-shoreability Coding

Training of coders. Westat conducted a one-day training session with 4 coders for this project. The training covered how to use the Standard Occupational Classification Manual 2000 (SOC) for coding occupations, the North American Industry Classification System 2002 (NAICS) for coding industries, the assigning of an offshoreability score to the occupation, and how to use the Microsoft Access system developed for the project. Much of the training was devoted to classification principals for SOC and NAICS coding. For off-shoreability, the coding scheme and job characteristics that need to be considered in applying a score were discussed in-depth. A summary of these job characteristics was included in the training materials and also on a separate reference sheet that coders could refer to throughout the project. The remaining portion of the training consisted of exercises where coders examined raw data responses from the questions of interest from the survey instrument and assigned SOC, offshoreability, and NAICS codes to them. The codes and the basis for assigning these codes were then discussed as a group.

Access tool for coding. Using Microsoft Access, Westat designed a tool for use by the coders in reading the raw data responses and assign a NAICS code, a SOC code, and an offshoreability code from the questions of interest from the CATI interview. The responses to questions Q4a or Q4b, Q5, and Q5a were used for assigning the NAICS code. The responses to questions Q6a or Q6b, Q7, and Q8 were used for assigning the SOC code and the offshoreability code. However, responses to all of the questions listed above were available to the coders no matter which code they were assigning, because

information relevant for assigning a code might be found in any of the responses. Once logging into the system, the tool presented six columns showing the raw data responses from the 6 questions noted above. For assigning the SOC and offshoreability codes the responses to questions Q4a or Q4b, Q5, and Q5a were presented to the left and the responses to questions Q6a or Q6b, Q7, and Q8 were off to the right. For assigning the NAICS code the order of the responses was reversed with the responses to questions Q6a or Q6b, Q7, and Q8 presented on the left and the responses to questions Q4a or Q4b, Q5, and Q5a were off to the right. If the coder was assigning a SOC and off-shoreability code, two additional columns consisted of tools for entering these values. The box for coding the offshoreability code was a drop-down box that listed the six offshoreability coding options (the 5-point off-shoreability scale plus a 0 code option that meant insufficient information or too vague for assigning a code) and their definitions. If the coder was assigning the NAICS code, one additional box was provided for entry of the NAICS code. A final column for all NAICS, SOC and off-shoreability coding presented a box for coders to enter a flag to indicate supervisor review or assistance is needed, or to enter an update flag for coding corrections.

The off-shoreability scores that had previously been assigned by Westat to SOC categories were also integrated into the coding too. For cases where a coder's assigned off-shoreability score did not match the score assigned to the SOC category, a report was generated for review of the case by the Westat supervisor.

Coding procedures. Each of the four coders received a randomly assigned one-fourth of the completed interviews. The coders examined the raw responses presented by the Access tool and assigned the SOC, off-shoreability, and NAICS codes. The SOC and off-shoreability coding were completed as a unit for each case and the coders were instructed to assign both codes before proceeding to the next case. The NAICS coding was done separately from the SOC and off-shoreability coding, although all responses from the questions of interest from the survey instrument were available for both types of coding, SOC and NAICS. The supervisor 100% verified the coding on an on-going basis throughout the coding process, providing close oversight and individual feedback about the coders' accuracy of assigning the NAICS, SOC, and off-shoreability codes. During the coding period the supervisor remained readily available to answer questions and provide guidance as needed.

Verification of Reported Occupational Licenses

Westat staff attempted to verify a randomly selected $1/12^{th}$ of all reported occupational licenses. The first step in this process involved asking for the full name of the respondent, along with the state (or city or county) in which the license was applicable. We collected up to three states/cities/counties in which the license was said to applicable. However, almost one-quarter of the respondents who reported an occupational license refused to provide their name to us – we were unable to proceed further in the verification of these cases. The next step involved searching for a suitable database on the internet by which we could verify that the respondent currently holds a

valid license. This generally meant examining state government websites, many (if not all) of which provide readily accessible internet-based tools for consumers and employers to verify that a given individual has a professional license. However, very few cities or counties appear to provide such tools. We also used internet search tools (google.com) extensively in an effort to find relevant information. For example, we performed searches on the individual name, paring it with additional information such as the occupation, the state or city, and the term "license."