Detailed Project Description

This year's project will study how to account for population members without internet access when conducting internet surveys. The internet is an attractive mode for conducting surveys because of the low cost of administering the survey and the elimination of interviewer effects that may cause bias in telephone and face-to-face surveys. It is very practical for surveying populations for which a complete list of email addresses is available. For example, a human resources department has a list of email addresses for all employees of their company, and a university registrar has a list of email addresses for all students enrolled in the university. However, when the target population is more general, such as parents of students enrolled at a particular elementary school or likely voters in an upcoming primary election, a list of email addresses is unlikely to be available, and many members of the target population may not have access to the internet at all. Mediamark Research estimated that in the spring of 2006, 65% of adults in the United States had internet access at home and 81% had access at home, work, or another location such as a public library. An internet survey of a population without universal internet access suffers from undercoverage because many members of the population are not a part of the sampling frame. Coverage bias may result if those excluded from the sampling frame differ from those who can be reached electronically on the variable(s) of interest. For example, if we want to survey parents regarding the amount of time their children spend engaged in watching TV, playing outside, playing computer or video games, and working on a computer, it is likely that the responses will differ for families who have and do not have internet access.

The weighting schemes already employed in most surveys can reduce the effects of coverage bias. Post-stratification adjusts the weights assigned to each observation so that the weighted proportion of observations that fall into a particular demographic category matches the proportion in the population that fall into that category. For example, the school survey might use neighborhood as a post-stratification variable. An internet survey would get relatively fewer responses from a low-income neighborhood, but the observations that are obtained from the low-income neighborhood would be weighted more heavily. If the families in the low-income neighborhood who do have internet access are more representative of the other families in their neighborhood than other families with internet access outside the neighborhood, then post-stratification reduces the coverage bias. However, depending upon the population, the variable of interest, and the post-stratification variables, post-stratification may or may not reduce the bias as much as survey administrators would like.

Commercial organizations, including Harris Interactive and Knowledge Networks, have been using the internet to conduct large-scale national polling for many years. Both have a large panel of respondents from which they draw samples. Knowledge Networks skirts the coverage problem by providing WebTV units to panel members and paying for the service. Harris uses a weighting method known as propensity scoring to account for undercoverage. They poll a separate telephone sample and weight individuals in the internet sample based upon how similar they are to those in the telephone poll. Many

organizations, however, do not have the means to conduct the telephone interviews required to make these weighting adjustments.

The novel question this project will address is whether or not any additional weighting to account for coverage bias, beyond post-stratification, can be done in a survey conducted strictly via the internet. This question is analogous to asking whether a survey conducted strictly via telephone can account for the coverage bias due to excluding households without telephone service. Duncan and Stasny (2001) addressed this issue using a telephone poll that asked respondents if they had been without phone service at any time in the last year, and if so, for how long they were without service. Respondents with an interruption in service were called transients and deemed to be more similar to households currently without phone service and therefore unreachable for this survey. Several weighting schemes involving transient status were analyzed, and some resulted in modest but significant differences in the estimates of variables of interest, without greatly increasing the variances of these estimates. A weakness of the research was the inability to test the weighting schemes against values unaffected by coverage bias since no surveys were conducted with non-phone households.

The data resource for this project will be the Current Population Survey's computer and internet supplement of October 2003. The Current Population Survey (CPS), conducted by the United States Census Bureau for the Bureau of Labor Statistics, is a monthly survey of approximately 50,000 households which provides information about the labor force characteristics of the U.S. population. While the supplement collects data about computer and internet usage, it is conducted through personal and telephone interviews and thus does not suffer from undercoverage. Using the subset of respondents with internet access as a proxy for a sample obtained via an internet survey will allow us to study the proposed weighting methods and their impact on estimates. The computer and internet supplement does not ask about interruptions in internet access, so we do not have an exact analogy to transient phone status. Even if a question about interruptions in internet access were to be included on a survey, transient internet access differs from transient phone status in that you do need to have phone service at the exact time an interviewer attempts to make contact in order to be eligible for a telephone sample while an email request to complete a survey may be stored on a server and accessed at a later date when internet access is restored. In this project, we plan to use respondents who do not have internet access at home, but who do use the internet at work or elsewhere to serve as the "transients", better representing the group without any internet access. We will employ several different weighting schemes which give more weight to the responses of the transients in the sample and study the effects of these schemes on point estimates. A benefit of this data set is that we will be able to easily measure improvements in accuracy due to our weighting schemes. The full data set of all households will provide the benchmark estimates against which we will compare our weighted estimates computed using only households with internet access.