

**Method and Representation in Internet-Based Survey Tools:
Mobility, Community, and Cultural Identity in Survey2000**

James C. Witte (Principal Investigator), Lisa M. Amoroso, and Philip E. N. Howard
Department of Sociology
Northwestern University

Tuesday, November 16, 1999

Abstract

The Survey2000 Project is the largest and most comprehensive Internet-based social science survey to date. Along with generating interesting data about geographic mobility, feelings of community, and culinary, literary and musical tastes, the experience of operating a survey with Internet tools has set into sharp relief important methodological issues of sample size, representation, and generalization. We argue that Internet-based survey research can yield meaningfully comparable data about both Internet users and larger populations.

KEYWORD LIST: Survey2000; Web-Based Survey Design; Sampling

Survey2000 is a collaborative research project of the staff at National Geographic Interactive and academic researchers. The National Geographic Society and Northwestern University have provided financial support for this project. One year after data collection is completed a public use version of the data set will be distributed to allow the academic and scientific community at large access to the data. Any information that could potentially compromise the anonymity of respondents will be eliminated from the public use data set. For their advice and assistance in the preparation of various drafts we are grateful to Bill Bainbridge, Bonnie Erickson, Joe Germuska, Wendy Griswold, Keith Hampton, Valerie May, Pete Peterson, and Barry Wellman.

Draft: under review, please contact jwitte@clemsun.edu regarding citation and publication status

Method and Representation in Internet-Based Survey Tools: Mobility, Community, and Cultural Identity in Survey2000

Tuesday, November 16, 1999

Abstract

The Survey2000 Project is the largest and most comprehensive Internet-based social science survey to date. Along with generating interesting data about geographic mobility, feelings of community, and culinary, literary and musical tastes, the experience of operating a survey with Internet tools has set into sharp relief important methodological issues of sample size, representation, and generalization. We argue that Internet-based survey research can yield meaningfully comparable data about both Internet users and larger populations.

KEYWORD LIST: Survey2000; Web-Based Survey Design; Sampling

INTRODUCTION

During two months in 1998, over 80,000 people collectively spent over 2 million minutes on-line at the National Geographic Society's official web site participating in an interactive first-of-its-kind survey on mobility, community, and cultural identity.¹ In this paper, we describe the project background, project development, and details of the survey content and technical design. This paper also addresses the critical methodological issues of sample representativeness and bias that arise with a voluntaristic Internet-based survey.

Survey2000 was on-line at www.nationalgeographic.com, the official web site of the National Geographic Society (NGS), during September and November of 1998. This project represents an unprecedented effort to use the rapidly growing power of the web to collect serious social science data. Survey2000 is a collaborative research project of the staff at National Geographic Interactive and academic researchers, funded by the National Geographic Society and Northwestern University. The survey focuses on geographic mobility, community, and cultural identity. A public use version of the data set will be available on CD-ROM to allow the academic and scientific community at large access to the data. Any information that could

potentially compromise the anonymity of respondents has been eliminated from the public use data set.

Different survey instruments are used for adult Canadian and US respondents, adult respondents from other countries and for children under the age of sixteen. Data collected in Survey2000 falls into several clusters: a) migration histories, b) measures of sense of community and c) measures of cultural values and tastes in food, music, and literature. This data can be used to address important research questions including:

- 1) How have the lifetime migration histories of individuals changed over time? Large-scale surveys have looked at whether individuals have moved over the past five or ten years. Other studies (most notably by the National Center for Education Statistics and the Bureau of Labor Statistics) have looked at lifetime mobility histories for particular cohorts. Survey2000 allows researchers to study changes in migration cohorts across generations.
- 2) Where do individuals in today's society find a sense of community? Geographic communities, extended family, voluntary associations, workplace relations, telephone and computer connections are all part of the mix. But how are these different components blended and how does the balance change according to individual characteristics?
- 3) To what degree do individuals in the US, but also around the world, share common values and tastes in food, music, and literature. How are these tastes associated with demographic characteristics, including the extent to which an individual has been geographically mobile? To what extent have regional tastes in music and literature declined only to be replaced by a modulated and standardized "McCulture"?

Survey2000 also represents an experiment in survey methodology, and experiment in an area marked by great potential but also little experience (Blank, 1997; Schaefer and Dillman, 1998; Fisher et al., 1996). From this experiment we learn:

- 4) To what degree can a web-based survey replace traditional survey research methods? How successful can different methods be in addressing the sampling issue? Also, can broad-based promotion and outreach efforts extend survey coverage to the general population?
- 5) Previous experiments with computer-assisted, personal or telephone interviewing (CAPI or CATI) systems are known to have improved data quality while allowing for more complex, individually tailored interview schedules. Survey2000 builds on this strength, but then adds the interactive potential of a hyper-media interface. The goal was to produce an instrument that is not only complex and customized, but also engaging so as to minimize respondent burden and respondent attrition.

PROJECT DEVELOPMENT

Eighteen months of collaboration are behind the Survey2000 instruments. The original impetus for the project came from staff at National Geographic Online with the idea that a web survey on the topic of population and migration would add to the NGS's coverage of modern society and millennial transition. From a survey research perspective, the nonrandom nature of a web survey sample raises serious questions; however this issues is not unique to web-based survey research and is likely to decline in significance as the web penetrates further into society (Smith 1997). The response to this challenge is detailed below in the section on sampling issues. In short, our solution depends on two mechanisms: 1) the survey relies on items from existing surveys conducted with traditional sampling and survey methods (e.g., the General Social Survey) to ensure that there are external benchmarks to assess the nature of the survey bias and construct the necessary weights, 2) extensive use of NGS public relations and community outreach resources to extend survey coverage.

The project decided to focus on geographic mobility as an independent variable and on individual values and preferences as dependent variables. Specifics of the survey instrument, however, were worked out by project collaborators. As the number of researchers interested in the data increased, so too did the commitment of the NGS. Indeed, the efforts of the academic collaborators and NGS were necessary complements to one another: academic interest in the project was necessary to build NGS support, while this support was needed to ensure the quality of the data, particularly with regard to the size and breadth of the sample.

SURVEY CONTENT

As noted above, Survey2000 consists of three main instruments: 1) the Canadian and US Adult Survey, 2) the Youth Survey, and 3) the International Respondent Survey. Upon connecting to the Survey2000 site, respondents are asked to choose the appropriate form. As an

incentive for completion, respondents are also told that a random number of participants will be awarded a gift from NGS. A follow-up screen queries each respondent's age and current citizenship; respondents who have mistakenly chosen the wrong survey form are reassigned based on this information. The major sections and key questions of each instrument are discussed below.

Canadian and US Adult Survey

The Canadian and US Adult Survey is the most complex and detailed of the three Survey2000 instruments. Respondents begin by supplying basic **demographic information**, including current primary residence, zip or postal code, marital status and household composition. Respondents are also asked to identify languages regularly spoken in the household; however, the survey is only presented in English. Further questions ask about race and ethnicity, educational enrollment and attainment and current employment status. Separate response codes are offered to US and Canadian respondents. Questions concerning race and ethnicity are worded to prompt the respondents to self-identify as they normally do on government forms. During the survey development phase we discussed providing a greater range of response categories (including open-ended fields as used in the 1980 US census). However, we decided that the benefits of comparability with external benchmarks were greater than the richness offered by a wider range of response categories. Open-ended responses are solicited regarding current occupation and most recent occupation for those persons not currently employed.

Questions concerning **Internet access and use** constitute a second important survey section. Respondents are asked where they are completing the survey (home, work, community center, or library) and how long they have used the Internet.² Respondents are also queried as to

the frequency with which they engage in specific Internet activities (with email, purchasing products, use of listservs).

The next block of questions concerns the respondents' individual **mobility history**. Respondents are asked if they have ever lived outside the US and Canada, how long they have lived at their current address and whether or not they have always lived within thirty miles (50 kilometers) of their current address. Subsequent questions ask about the number of different dwelling units occupied, about other members of their current household, and whether or not other relatives currently live in the immediate area.

The mobility history questions then solicit respondents' place of birth. First, respondents are asked to pick from a list of geographic landmasses (North America, South America, Europe, Asia, Africa, and Oceania), followed by a list of countries associated with each. Respondents who were born in the US or Canada are then sent to a further screen that asks the respondent to choose the state or province of their birth. Based on this choice, the respondent is asked to select the closest location from a list of ten to fifteen cities in that state or province. At this point respondents are asked a similar sequence of questions regarding their residence at ages seven, fourteen, twenty-one, twenty-eight, thirty-five, forty-two, fifty, sixty, seventy, eighty and ninety. To shorten this block of questions, a filter question first determines at each age, if there has been no change in residence since the previous age. Previous responses regarding duration at current location and birth year are also used to minimize respondent burden and to avoid asking for information that can be obtained from previous replies. Internal checks helped verify that respondents were consistent in the information they provided. When respondents were inconsistent, they were asked for clarification.

Information detailing each respondent's **social world** is collected during the next block of questions. Respondents are asked how often and in what way (i.e., personal visits, phone calls, faxes, letters, cards or email) they have social contact with 1) relatives who live less than 30 miles away? 2) friends who live less than 30 miles away? 3) relatives who live more than 30 miles away? and 4) friends who live more than 30 miles away? A separate question then asks for the frequency with which the respondent gives or receives help or assistance from these same four sets of relatives and friends. Next, building on a series of questions regularly included in the General Social Survey (GSS) respondents are asked about their membership in a set of formal organizations (e.g., service clubs, veterans groups, labor unions and social advocacy groups). In addition, respondents are asked about various forms of political involvement (e.g., social groups). This section then closes with a series of Likert-scale (seven points ranging between strongly agree and strongly disagree) items related to community. These items³ pertain to traditional sources of community as well as Internet-based communities.

This section concludes with a series of questions about the recreational and leisure time activities that are an important part of the context of an individual's social world. Survey2000 builds on the GSS questions about involvement in a range of activities, such as gardening, reading, sports and adds additional categories, (e.g., renting videos, going to casinos and attending work-related social events) that round out the list. This section concludes with two items that assess the extent of knowledge and interest a respondent has in music, literature and food, the three areas of cultural identity that form the basis for the concluding section of the survey.

The final section of Survey2000 is titled **interests and perspectives**. Each respondent is presented with one of four randomly selected topical modules. The four topical modules are

literature, food, music, and views on the world. The literature and food modules are set up quite similarly. In each module a customized list of items—authors in the case of literature and dishes in the case of food—are constructed for each respondent. These lists include twenty-eight items that represent the geographic area of residence at selected ages, items representing areas where the respondent never lived, and items presumed to transcend any particular region. Respondents are asked to indicate their degree of familiarity and preference for each author or dish.

Respondents who receive the music module are presented with a list of music genres and asked to indicate their familiarity and preference for each type; this list is identical for all respondents. Each respondent then assesses a smaller set of genres, in an effort to more precisely understand his or her knowledge of the variation within a given genre. In some cases (when classical, jazz, country or dance music are presented), respondents are offered the chance to hear sound clips that represent particular sub-genres. The views-on-the-world topical module consists of eighteen Likert-scale items that complement the community questions asked of all respondents. These items tap respondents' views of the world (e.g., complexity, optimism, and altruism) and include a subset of Internet-related items.⁴

Though each respondent initially receives a single topical module, they are also offered the option to complete the remaining topical modules and to comment on two open-ended questions. The survey closes with a screen that informs a randomly selected subgroup of the sample that they have been selected to receive a gift for participating in the survey. As a further incentive, all respondents are offered a customized set of web links (URLs) that have been selected to reflect their individual responses to questionnaire items, including geographic areas they have lived in and their leisure time activities and interests.

International Respondent Survey

This survey instrument is designed for adult respondents who are neither US nor Canadian residents or citizens. In theory, it is easy to imagine a survey instrument for all respondents with a similar structure as that used for US and Canadian respondents. However, early in the project development process it was decided that the Survey2000 lacked the resources necessary to develop parallel instruments for all international respondents. Moreover, the organizational apparatus of the NGS, which is seen as a direct means to counter the bias inherent in a web survey, that is central to the US and Canadian adult survey, loses some of its efficacy outside North America.

Nonetheless, Survey2000 is a *worldwide* web survey. Defined in this manner, Survey2000 could hardly ignore respondents from outside North America or force them to respond to an instrument that showed little sensitivity to a significant subset of respondents. Thus, the decision was made to turn a potential limitation of Survey2000 into an advantage. Survey2000 offers a unique opportunity to empirically test of the claim that American cultural imperialism has cause the emergence of a global McCulture, at the expense of regional and national cultural diversity. Survey2000 over-sampled well-educated and well-off respondents and this bias is likely to be more acute among international respondents. Furthermore, the most well educated and well-off segments of the world's population are most likely to have been exposed to and adopted cultural hallmarks of North America. To the extent that North American culture has left a hegemonic footprint, it should therefore be particularly apparent among Survey2000 respondents.

To consider this question the international respondent instrument begins with a standard set of demographic questions, residence, citizenship, age and gender, as well as marital status, household composition, educational attainment, and employment status. The international

respondents also receive the same set of Internet use and access questions as North American respondents. The mobility history for international respondents is limited to current residence and place of birth. In addition, respondents are asked if they have ever visited or lived in the US or any other country than their current country of residence. International respondents receive the same set of questions regarding their social world, including contact with friends and relatives, group membership, political participation, and the Likert-scale community items.

The interests and perspectives section of the international survey is necessarily quite different from the US and Canadian instrument, which customizes the literature and food modules according to each respondent's personal geographic mobility history.⁵ The international survey does not randomly allocate respondents to a single topical module, but instead gives each respondent an abbreviated version of the literature, food, and music modules. Each international respondent is provided a list of eight North American authors and eight North American dishes drawn from a pool of items presumed to transcend North American regional culture. The respondents receive the music topical module in its entirety. These items clearly do not reflect the diversity of world culture, though this design is well suited to mapping the spread of North American culture among a certain segment of the world.

Youth Survey

All respondents under the age of sixteen, regardless of nationality, are directed to the youth survey. The survey content differs substantially depending on whether the respondent is between the ages of thirteen and fifteen or twelve and younger. All children are asked to request parental permission to complete the survey, and then queried as to gender, citizenship, current residence (including zip or postal code), where they are completing the survey (e.g., home, school, parents workplace) and languages regularly spoken at home. Youth respondents are also

presented with an abbreviated version of the mobility history with questions focusing on length of residence in current location, total number of residences occupied and place of birth. The social world set of questions for youth respondents focuses on household composition and activities undertaken with parents and guardians. Children aged thirteen through fifteen are also asked about parental supervision, peer values, parental involvement in school activities and neighborhood safety and solidarity. Youth respondents of all ages are given standard self-esteem items, which come from the child supplement to the Panel Study of Income Dynamics (PSID) for the younger children and from the National Longitudinal Study of Youth (NLSY79) for those age thirteen through fifteen. The older youth respondents' items also include measures of locus of control and propensities toward risk-taking. Finally, the older youth respondents also receive a short set of items designed to measure attitudes toward the future, with a special emphasis on the next millennium.

THE WEB-BASED SURVEY DESIGN

As survey research has become increasingly sophisticated, social scientists, in particular sociologists, have become increasingly aware of the extent to which the findings of survey research are part and parcel of the social process and technology of data collection.⁶ Survey research requires social interaction, which in turn is sensitive to the technology upon which the process rests. Different survey research techniques face-to-face interviews, self-administered paper and pencil questionnaires, telephone interviews, CAPI, and CATI not only depend on different technologies, but also organize the social dynamics of data collection in a different fashion (Bratton and Newsted, 1995). Thus, a basic understanding of the technology behind a Web-based survey, such as Survey2000, is crucial to understanding the overall dynamics of this new means to collect social science data (Kehoe and Pitkow, 1996).⁷

Program features

Survey2000 is delivered to the respondent population via a PERL script. The script takes each page as it is submitted, writes the data into an Oracle database, and determines which page should be presented next. Wherever possible, the script suppresses questions that have been implicitly answered or rendered irrelevant by earlier responses.

In the case of food and literature questions, using a web-based script supports flexibility that would never be possible in a self-administered paper survey. The lists of dishes and authors delivered to each respondent are custom-built based on the various locations the respondent reported living in during an earlier segment of the survey. PERL and Oracle were used for this project mostly as matters of convenience and familiarity. Because the data is in a conventional flat-file format, there is no application for Oracle's relational functionality.

One benefit of a customized survey approach is that it shortens survey time. For example, questions concerning the frequency with which an individual engages in specific Internet activities is not asked of those respondents who say that they are using the Internet for the first time to complete Survey2000. Similarly, respondents who say that they live alone are not queried as to whether specific relatives live in their household. The greatest efficiency gains are realized in the geographic mobility section.

Design elements

The design elements of Survey2000 are an essential feature of the project. Most fundamentally, the project's affiliation with the National Geographic Society is intended to provide the project with a credibility and a sampling platform that few Web sites can offer. The NGS web site is well designed, regularly maintained, and attracts approximately 1.5 million "hits" per month. During the two-month period of data collection, a link to the survey was

placed on the NGS home page. References to the survey site were also published in the NGS's adult and children periodicals as well.

The NGS web design group used its considerable experience to enhance the aesthetics and functionality of the survey layout. Once a respondent begins the survey, the NGS logo remains on every page; however, there are no links to other pages or other sites to tempt the respondent into going elsewhere. The goal is not simply to capture respondents, but to engage and reward them for their participation. For example, the sidebar on the mobility history screens is customized for each respondent. When a respondent is asked where she lived at birth or at a given year the sidebar text reminds the respondent of the year in question and list three events that took place in that year.⁸ These facts are designed as prompts, but also as a way to make the process a bit more entertaining. Moreover, beyond the programming logic described above, an effort is made in the survey design to reduce respondent burden as well. For example, “checkboxes” and “radio buttons” are used extensively to minimize the respondents' keyboard input.

Pre-test results revealed that respondent burden might be too heavy if each respondent were expected to complete all four of the cultural modules. Thus, each respondent was given one of the four cultural modules; after completing the base survey and one of the four topical modules, each respondent received a thank you page, that included the option to continue the survey and respond to the three remaining topical modules. This approach might have had cost of its own with respect to possibilities for analysis across cultural domains; however, over half of the US adults (N=23,384) voluntarily continued after completing the base module. Thus, researchers can consider correlations across cultural domains – are individuals' preferences in music and food driven by the same factors that affect their literary tastes?

SAMPLE OVERVIEW

Tables 1A and 1B provide an overview of the entire Survey2000 sample and their form of participation. Over 80,000 surveys were initiated and just over 50,000 were completed. Adults living in or citizens of the US or Canada initiated the most surveys (N=45,951) and completed over 37,000 surveys. Adults from other part of the world comprised about one-fifth of the initiated surveys. Youth surveys were completed by 9,785 children in the US, 970 Canadian children, and 1,635 international youth. The overall survey completion rate was over 70% for all adults and almost 60% for children.

TABLES 1A and 1B HERE

Just about half of the initiated surveys (40,612) were from US adult respondents. Combined with an 80.5% completion rate, this yields a sample of 32,688 complete US adult surveys. For many of the questions these data address, partial responses are interest. Thus, a sample size of more than 32,688 respondents exists for relevant demographic and social capital measures that were the start of the survey instrument.

As noted above, for many applications, the critical issue concerning Survey2000 is the extent to which we can make generalizations from this sample to larger populations. The results presented in Table 2, which compares the Survey2000 sample with recent GSS surveys and US Census Bureau statistics concerning central demographic variables, ought to be viewed in two ways. First, these results indicate the extent to which the Survey2000 results will be statistically adjusted to adequately represent the US population and the Internet population within the US. Second, the Survey2000 sample provides some insight into the magnitude of the difference between the general US population and its Internet population.

TABLE 2 HERE

Recent findings indicate that the gender gap in Internet use for U.S. adults has essentially disappeared, and trends towards equal access by education, income, and race have also been noted (Media Metrix, WIRED, July 1999; Clemente 1998; Katz 1997). Particularly concerning gender, our results corroborate these findings. This comparison shows, for example, that while just over half (50.7%) of the Survey2000 sample is male, female respondents constitute the majority in the 1996 (55.7%) and 1993 (57.2%) GSS samples.⁹ Further, the Survey2000 sample is considerably younger, with a median age of 38 years than that estimated by the GSS (44 years in 1996 and 43 years in 1993).

The Survey2000 sample supports the widely held view that minorities are under-represented on the Internet: 92.5% of the respondents are white.¹⁰ Only 1.5% of the US adult surveys are from African-Americans. In subsequent efforts to generalize from Survey2000 to the broader US population weights will be developed to make the necessary statistical adjustments. But it should also be pointed out that the large sample size should make this possible. Though only 1.5% of the respondents are African-American, this amounts to 538 surveys. The 1993 GSS only has 179 African Americans, while the GSS African-American over samples in 1982 and 1987 include 354 and 353 African-American respondents respectively. Even in 1996, when the GSS sample was doubled, African-Americans only number 402. Thus, the actual proportions across categories are not as important as the number of respondents within each cell. With the large number of African-Americans who completed Survey2000 we will be able to draw inferences using statistical theory.

Finally, Table 2 indicates a large difference between the educational makeup of the Survey2000 sample and the US population at-large. Only nine-tenths of a percent of Survey2000 respondents have less than a high school degree, as compared to 15.2% of the 1996

GSS sample and 18.1% of the 1998 GSS sample. The GSS estimates are quite consistent with Census Bureau 1998 statistics that indicate that 17.9% of the US population age 18 or older has less than a high school degree. Also, the proportion of Survey2000 respondents with a high school degree but no post-secondary degree (31.9%) is considerably lower than that found in the 1996 GSS (54.1%) and the 1993 GSS (52.5%). Correspondingly, respondents with post-secondary degrees are over represented in the Survey2000 sample. The proportion of Survey2000 respondents with an Associate's degree is quite similar to the Census Bureau statistics; but the proportion with a Bachelor's degree (34.1%) is roughly double that provided by the Census Bureau. The proportion of Survey2000 respondents with a graduate degree (25.2%) is more than three times the official Census population estimates. Once again, this means that weighting will be required to make any generalizations from the sample to the US population at large, but in and of themselves these numbers reveal a great deal about the educational background of the Internet community. These data are instrumental in assisting researchers' early attempts to represent the population of Internet users.

SAMPLING ISSUES: RANDOMNESS AND REPRESENTATION

The potential for sample bias in data collected from the Internet represented the most serious methodological facing Survey2000. Critics of the project are quick to recall the famous Literary Digest poll that predicted Landon's victory over Roosevelt. This poll had a sample size more than 2 million, but still came to the wrong conclusion. There are some superficial similarities between Survey2000 and the Literary Digest Poll, but there are significant differences as well. The Literary Digest poll made no effort to assess the representativeness of its sample, while the Survey2000 explicitly incorporates features designed to measure selection bias and to compensate for the fact that the sample will not be random.

The goal of survey research is to collect data on a sample that represents a population. Randomness does not guarantee representativeness; rather it provides the means to quantify the level of confidence with which one can say that the sample represents the population. In a simple random sample, all members of the population have an equal and known probability of being selected into the sample. In practice, this assumption is rarely met. The poor and the rich are likely to be underrepresented in telephone surveys and the homeless are undercounted in samples based on dwelling units. In other cases, a sample may be designed so that the selection probability varies between different strata of the population. For example, minorities are routinely over sampled to increase the absolute number of minority sample members. Deviations from simple random sampling in terms of unequal selection probabilities are of little statistical concern, so long as the probability of selection is known. Departures from equal probability sampling are routinely handled through weighting procedures, which take into account the differences between individuals in sample selection probabilities and deflate or inflate the observed outcomes accordingly.

Greater difficulty arises, however, when as in the case of Survey2000, the probabilities of sample selection are unknown. In other words, this is a study of a hidden population where the size and boundaries of membership are unknown. Snowball sampling is a widely accepted approach to studying hidden populations, and is essentially the approach taken by Survey2000. The survey may not yield a random sample and we do not know the selection probabilities for sample members. This does not mean that the survey can not yield representative social science data. To begin with, though we do not know the selection probabilities our data allow us to estimate these probabilities. The survey collects data on standard demographic characteristics (e.g., gender, age, race, education, etc.) and combinations of these attributes for the sample can

be compared to official government statistics. The selection bias is also likely to be correlated with other factors — such as attitudes and values toward community and culture — that cut across standard demographic variables. For this reason, a number of items used in Survey2000 are based on other studies including the GSS, PSID, and the NLSY79. These studies are based on traditional sampling and data collection methods. These items serve as external benchmarks.

Based on these and additional benchmarks, it is possible to estimate the selection probabilities and construct adjustment factors, treating the sample as if it were random. This point may be illustrated with a simple example. Imagine a telephone survey of 1,000 respondents conducted during the daytime. If the population is split equally between men and women, then men and women have an equal probability of being in the sample. However, after conducting the survey we find a sample with 400 men and 600 women (despite increased female labor force participation, women are more likely to be at home and answer the phone during the day). The selection probability for men is 0.8 ($400/500$) and for women is 1.2 ($600/500$). Multiplying the observed sample of 400 men by 1.25 (the inverse of 0.8) and the observed sample of 600 women by 0.8333 (the inverse of 1.2) yields a weighted sample of 500 men and 500 women.

Moreover, just because the selection probabilities are unknown, the estimated selection probabilities do not necessarily vary greatly from the true (unobserved) selection probabilities. Since the selection probabilities are estimated, however, careful attention must be paid to the stability of these estimates and to the robustness of the weighted results. A sensitivity analysis is a simple and effective way to measure this: differing selection probabilities can be estimated and used to examine the extent to which the interpretation of the data varies with the choice of selection probabilities.

The issue of representativeness also raises the question of sample size, in particular how many respondents are needed to obtain a representative sample. Survey2000 is not based on the naive view that the bigger the sample, the better the sample. To begin with, the relationship between sample size and the precision of one's inferences is not linear; there are diminishing marginal returns to sample size. Furthermore, no matter how large the sample, size never guarantees representativeness. In fact, a sample of any size may be representative. Focus groups often include well under a dozen members, while a single key informant may accurately represent an entire group. The advantage of a random sample design (where the intent is to use a random sample and then quantify the probability with which one's sample does [not] represent the population) is that sufficient sample size may be determined more exactly.

With a random sample the optimal sample size is a function of several factors: the probability with which one is willing to reject a null hypothesis that is true, the probability with which one is willing to fail to reject a null hypothesis that is false, the degree of substantive difference between groups that one considers important and the within-group variance among members of the groups being compared. There should be no problem obtaining a large sample with a web-based survey of this type, at least no problem in terms of overall sample size. However, testing for differences in values between subgroups, the issue is not the overall sample size, but rather the sample size of each of the subgroups one wishes to compare. A common standard is that to compare subgroups a minimum of 30 respondents is needed *for each subgroup* (Fink, 1995: 43).¹¹ For certain subgroups—for example, educated, African-American females, over the age of 60, living in rural areas—one can imagine that it will take extraordinary efforts to survey an adequate number of respondents. Indeed this is one reason why the National Geographic Society's participation in this project is so important. The Society used other

National Geographic media—the magazine, television, and school-based educational activities—to encourage participation from a wider range of people than those normally found on the Web.

SAMPLING ISSUES: A PRACTICAL WEB APPLICATION WITH SURVEY2000

Two elements of the Survey2000 design may allow sample bias. First, as with voluntary CAPI or CATI surveys, respondents may have different – and difficult to identify – attributes from respondents who would not have participated voluntarily. Second, the institutional character of the survey host organization may have limited the selection of people invited to participate. These possible biases deserve serious assessment, and we can evaluate any bias by comparing our sample to other samples of the kinds of population we hope to represent. Since we are interested in the ability of Survey2000 data to represent both the Internet population and the general US population, we compare our sample with kinds of samples of these two populations.

Voluntary Surveys

Similar to telephone, mail and face-to-face surveying, Survey2000 requires volunteers to offer information, and the respondent can end the interaction at any point. However, the essential anonymity of a web survey removes all elements of formal and informal social control that encourage telephone, postal, and face-to-face respondent cooperation. Although we were unable to assess the respondent's willingness to continue at any given moment, we used a creative and engaging survey design. As one respondent noted: "It was fun to take." Beyond this and other anecdotal evidence, our completion rates speak to the engaging nature of the design.

Migration histories and questions about community were given to all respondents as part of the base survey. To reduce respondent burden, each respondent was given one of four

modules measuring cultural values and tastes. After completing the base survey and one of the four cultural modules, each respondent received a thank you page that included the option to continue the survey and respond to the three remaining topical modules. Of the 40,642 adults who began Survey2000, 80% completed the base questions and one of the cultural modules, 58% went on to begin the three remaining cultural modules, and 50% completed the three extra cultural modules. In other words, almost three-quarters of those who completed the base questions and one cultural module volunteered to continue the survey. Moreover, we retained data from partially completed surveys allowing scholars to more accurately analyze patterns of attrition and item non-response.

Snowball Sample

Since much of the snowball sample was generated by the National Geographic Society, we acknowledge a possible sample bias in that Survey 2000 respondents will probably have many of the attributes of typical visitors to the National Geographic Society website. However, publicity was generated over listservs and through articles in several magazines and newspapers. For example, over a two-day period in which HotWired Magazine provided a direct link to the survey some 2600 surveys were initiated, though on average 430 surveys were initiated on each day of the life of the survey. Other newspaper articles and outreach efforts into libraries and schools also generated publicity.

From what we know of Internet culture through other anecdotes and surveys, people who responded to the NGS outreach effort are also likely to have many of the attributes we associate with typical Internet users: middle class, educated, either students still in school or retired etc. These survey results will provide a slightly conservative estimate of Internet culture and demography. Moreover, the National Geographic Society is about as ideologically neutral as a

large public organization can possibly be, and we would be more concerned about bias induced by the survey host if another more controversial organization had hosted and publicized the survey.

Generalizing to the population of Internet users.

To assess the quality of our sample, we can compare the distribution of Survey2000 respondents with the general distribution of Internet users across the United States. For example, there is close correspondence between the number of Survey2000 respondents and Internet Service Providers (ISPs) across the country. The correlation of .968 (significant at $\alpha < .01$) suggests that Survey2000 respondents were provided Internet services by a representative sample of companies across the United States. In other words, the distribution of survey respondents by state is similar to the distribution of Internet service providers by state¹². Even in the case of California, where 14% of the country's ISPs operate, Survey2000 obtained 12% of its respondents, leaving only a small difference in the ratio of respondents to providers (1 Californian respondent for every 1.17 Californian ISPs).

Generalizing to the US population.

A second issue regarding the Survey2000 sample concerns the extent to which it represents the population off-line. The results presented in Table 2 clearly indicate that particular demographic groups are strongly over-represented in the sample and others notably underrepresented. The central issue then becomes, whether or not subgroups of Survey2000 respondents represent subgroups within the general population. To explore the depth of representativeness, other items from Survey2000 can be compared with similar GSS items. For example, both the Survey2000 and GSS population of single, white males between the ages of 19 and 40 with at least a BA have similar musical tastes. The comparison is not perfect since most

importantly the GSS data was collected in 1993, five years earlier than Survey2000. Thus, shifting musical tastes among members of this subgroup confound the comparison of the two samples. Nonetheless, such a comparison provides a useful starting point in efforts to generalize from Survey2000 to the population at-large. If the responses closely coincide, particularly with respect to music genres that have not experienced large shifts in popularity, then a statistical adjustment process that develops weights based on demographic characteristics may be adequate. However, if large differences are found, then the any weighting scheme needs to consider cultural preferences and tastes to avoid unwarranted generalizations from the Survey2000 sample to the population at large.

Specifically, Survey2000 respondents were queried regarding twenty genres of music and sixteen of these closely correspond to categories used in the 1993 GSS. Table 3 presents comparisons using seven illustrative music genres for the GSS and Survey2000 sub-samples of white males, between the ages of 19 and 40, with a Bachelor's degree. Considering these respondents' reactions to the "Bigband/Swing" genre, a clear gap is evident: of the GSS respondents 7.9% "Like it very much" and another 48.3% "Like it", while 24.9% of the Survey2000 respondents "Like it very much" and 42.8% "Like it". However, "Bigband/swing" music has also gone through a noticeable increase in popularity between 1993 and 1998, particularly among younger adults. Thus, it is difficult to say how much of the difference is due to differences in the population each sample represents and how much is due to shifts in preferences for "Bigband/swing" music among educated, young, white males.

TABLE 3 HERE

Considering "Blues and R&B" the two samples are rather similar: 57.8% of the GSS respondents favorably respond to this genre, as compared with 60.1% of the Survey2000

respondents.¹³ Significantly larger differences between the two samples are found comparing feelings regarding classical music. Among GSS respondents 58.7% reacted favorably to classical music, while 82% of the Survey2000 respondents in this demographic group were positively disposed toward the genre. As there is no evidence of a large-scale shift in tastes toward classical music between 1993 and 1998, this difference suggests an important difference in the populations represented by the two samples. Indeed, looking across the remaining columns there are substantial differences between the two samples in their responses to specific genres: GSS respondents reacted more favorably to the “Contemporary Pop/Rock” and “Country-Western” than the Survey2000 respondents, who were more likely to respond positively to the “Jazz” and “Latin e.g., Mariachi, Salsa” genres.

Taken as a whole, these findings strongly indicate that simply adjusting the Survey2000 sample weights to the marginals for central demographic variables would not yield plausible generalizations to the population at large. On the other hand, preferences regarding specific music genres may provide the analytical leverage to construct plausible weights. Tastes in music may be taken as indicators of a broad range of cultural characteristics, weighting up those Survey2000 respondents with music preferences similar to the GSS results, while weighting down those respondents with dissimilar preferences should capture some of the unobserved selection differences between the two samples. While this approach is far from perfect, it does afford one means to further consider the extent to which Survey2000 can effectively represent the population at large.

In sum, the design for Survey2000 is not based strictly on the principles of random sampling, which permit one to exactly know the probability that observed differences in the sample represent real differences in the population. Rather, a large snowball sampling approach

along with external benchmarks was employed. Survey2000 guarantees fascinating insights into the demography and sociology of the population of Web users. There has been little empirical research in large-scale web survey administration. These data provide a solid foundation upon which to build future projects.

SURVEY2000: LOOKING TO THE FUTURE

Survey2000-2, “Measuring and Maintaining Biological and Social Diversity.” is currently being planned. This second instrument will also be hosted by the NGS web site and is scheduled to go online in the fall of the year 2000. As with the earlier effort, the second data collection effort is designed to further our knowledge about the web as a survey research tool, while collecting data of broad topical interest. This effort will include a parallel phone survey.

In the short time this project has been underway, developments in Internet technology have only increased the viability of similar projects. As the wiring of the world expands, sample coverage becomes less problematic; as so-called “Internet push-technology” develops, the possibility of Internet-based probability sampling draws nearer. The development of Internet II and related technologies foreshadow an era when clickable maps to record respondent geographic mobility and widespread use of other multi-media survey tools will be commonplace. Finally, coming to grips with new tools for survey research should bring a new sensitivity to survey research as a process of social interaction.

The exercise of mounting such a large Internet-based survey project forced researchers to think carefully about issues of sample size, representation, and generalization. This was especially important since the Internet population, until Survey2000, had not been comprehensively surveyed. The results are being analyzed and interpreted by a diverse group of scholars, but these preliminary findings suggest that Survey2000 will substantially add to our

understanding of both the Internet population and of the practice of social science with contemporary research tools.

REFERENCES

- Blank, Grant. (1997). The Road Ahead: Observations on the Role of the Internet. *Social Science Computer Review*, 15 (2) Summer, 190-195.
- Bratton, Gregory R. and Peter R. Newsted. (1995). Response Effects and computer-administered questionnaires: the role of the entry task and previous computer experience. *Behavior and Information Technology*, 14 (5), 300-312.
- Clemente, Peter. (1998). *The State of the Net: The New Frontier*. McGraw Hill: NY.
- Fink, Arlene. (1995) *How to Sample in Surveys*. Sage Publications.
- Fisher, Bonnie, Michael Margolis and David Resnick. (1996, Spring). Breaking Ground on the Virtual Frontier: Surveying Civic Life on the Internet. *The American Sociologist*.
- Heckathorn, Douglas D. (1997) May. Respondent-Driven Sampling: A new approach to the study of hidden populations. *Social Problems*, 44 (2), 174-199.
- Katz, James. (1997, March/April). *The Social Side of Information Networking*. Society.
- Kehoe, Colleen M. and Jim Pitkow. (1996). Surveying the Territory: GVU's Five WWW User Surveys. *The World Wide Web Journal*, 1 (3), 77-84.
- Kiesler, Sara and Lee S. Sproull. (1986). Response Effects in the Electronic Survey. *Public Opinion Quarterly*, 50, 402-413.
- Pitkow, Jim and Margaret M. Recker. (1995). Using the Web as a Survey Tool: Results from the Second WWW User Survey, *Journal of Computer Networks and ISDN systems*, 27 (6).
- Schaefer, David R. and Don A. Dillman. (1998). Development of a Standard E-Mail Methodology. *Public Opinion Quarterly*, 62, 378-397.
- Schutz, Alfred. (1971). *Collected papers, Vol. 1*. Edited and introduced by Maurice Natanson. The Hague: M. Nijhoff.
- Sellen, Abigail. "Remote Conversations: The Effect of Mediating Talk with Technology" *Human-computer Interaction* 10, 1995, pp. 401-444.
- Smith, Christine B. (1997, June). Casting the Net: Surveying and Internet Population. *Journal of Computer-Mediated Communication*, 3 (1).
- Wired Magazine. (1999, July). The Condé Nast Publications, Inc.: San Francisco, CA.

ENDNOTES

¹ The Survey2000 world-wide web site can be visited at www.nationalgeographic.com.

² Respondents' Internet protocol (IP) addresses are also recorded as part of the Survey2000 process, as is the case in all Internet connections. This information can uniquely identify connected machines (though not respondents) if the machine has a static IP address. However, most respondents entered Survey2000 through public access Internet providers that assign the same IP address to a large number of respondents. This safeguards the identity of individual respondents, but at the same time permits the analysis of correlations among respondents associated with the same IP address. In any event, IP addresses will not be distributed as part of the public use data file.

³ These items include: 1) I feel close to other people in my community. 2) My daily activities do not create anything worthwhile for my community. 3) My community is a source of comfort. 4) I feel a sense of community with the people I've met on the Internet. 5) I have made new friends by meeting people on the Internet. 6) The Internet has brought my immediate family closer together. 7) The Internet has brought my extended family closer together.

⁴ 1) The world is too complex for me. 2) I don't feel I belong to anything I'd call a community. 3) People who do a favor expect nothing in return. 4) I have something valuable to give to the world. 5) The world is becoming a better place for everyone. 6) I cannot make sense of what's going on in the world. 7) Society has stopped making progress. 8) People do not care about other people's problems. 9) I find it easy to predict what will happen next in society. 10) Society isn't improving for people like me. 11) I believe that people are kind. 12) I have nothing important to contribute to society. 13) Talking with people on the Internet is as safe as communicating with people in other ways. 14) The Internet has allowed me to communicate with all kinds of interesting people I otherwise would never have interacted with. 15) The Internet isolates people from one another. 16) I feel I belong to an on-line community on the Internet. 17) Information on the Internet is as trustworthy as information from television and newspapers. 18) I can find people who share my exact interests more easily on the Internet than I can in my daily life off-line.

⁵ Not only is the mobility history for international respondents not as detailed as that collected for North American respondents, but also collecting the information on world literature and cuisine requisite to implement this design worldwide is beyond the scope of the Survey2000 project.

⁶ The significance of this relationship is noted quite clearly by Alfred Schutz (1971[1953]) in his discussion of how the social scientist's field of inquiry fundamentally differs from that of the natural scientist: "His observational field, the social world, is not essentially structureless. It has a particular meaning and relevance structure for the human beings living, thinking and acting therein."

⁷ Another interesting Internet survey project is hosted by the Graphics, Visualization, and Usability Center of the Georgia Institute of Technology, which has been conducting web-based surveys for 5 years. The GUVU's 10th survey offered cash incentives to respondents, advertised corporate sponsorship, and collected over 5000 responses. Whereas the GUVU is focussed on market penetration of Internet technologies and the rise of Internet usage, the Survey2000 is also focussed on how mobility shapes community values and cultural awareness. While GUVU data allows generalization about some features of the Internet population over time, the Survey2000 allows both generalization about the internet population over time and comparison with larger populations.

⁸ Two examples of the sidebars: For 1972: J. Edgar Hoover, controversial director of the U.S. Federal Bureau of Investigation dies. Arab terrorists massacre Israeli athletes at the XX Olympiad in Munich. U.S. first-class postage: 8 cents. For 1968: Martin Luther King, Jr., and Robert F. Kennedy are assassinated two months apart. Film director Stanley Kubrik releases 2001: A Space Odyssey. U.S. first-class postage: 6 cents.

⁹ A long standing issue with the GSS and other probability samples has been the overrepresentation of females (Smith 1979).

¹⁰ Another 713 respondents did not provide information concerning race. Presumably, many of these were not white. However, even if one assumes that all of those who failed to identify with one of the race categories are not white, still more than 90% of the sample is white.

¹¹ Thinking solely about important demographic attributes—gender, age (4 categories), marital status (3 categories) race/ethnicity (3 categories), educational achievement (3 categories), employment status (2 categories) and urban-suburban/rural residence (2 categories)—the number of unique combinations of attributes grows quickly. A sample of 25,920 would be necessary to compare each of these to any other combination. Smaller samples, however, are acceptable if similar subgroups are combined for a particular analyses. This is often the case and many "nationally representative samples," including GSS, are much smaller.

¹² We assume that competition among Internet providers will make the quality and quantity of ISPs proportional to

market demand throughout the country, and that roughly the same proportion of Internet users in each state have chosen to use regional or national Internet service providers.

¹³ Survey2000 respondents are somewhat more likely to say they “Like it very much” than GSS respondents.

However, given the magnitude of the difference (13.8% of GSS respondents as compared to 18.4% of Survey2000 respondents), this difference may also represent differences in survey method rather than differences in preferences for this genre, particularly given evidence that extreme responses are more likely in electronic surveys (Kiesler and Sproull, 1986).