

Research Methodology

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

JAMES C. WITTE

Clemson University

LISA M. AMOROSO

Northwestern University

PHILIP E. N. HOWARD

Northwestern University

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. The authors argue that Internet-based survey research can yield meaningfully comparable data about both Internet users and larger populations.

Keywords: Survey2000, web-based surveys, web-based survey design, sampling, Internet, World Wide Web, survey

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

Survey2000 was online at www.nationalgeographic.com, the official web site of the NGS, in the fall 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 and has been funded by the NGS, Northwestern University, and Clemson University. The survey focuses on geographic mobility, community, and cultural identity. A pub-

AUTHORS' NOTE: Survey2000 is a collaborative research project of the staff at National Geographic Interactive and academic researchers. The National Geographic Society, Northwestern University, and Clemson University have provided support for this project. 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, Barry Wellman, and Nathan Wright.

Social Science Computer Review, Vol. 18 No. 2, Summer 2000 179-195
© 2000 Sage Publications, Inc.

lic use version of the data set is available.² 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 U.S. respondents, adult respondents from other countries, and children under the age of 16. Data collected in Survey2000 falls into several clusters: (a) respondents' demographic characteristics, including the extent and duration of their Internet experience; (b) migration histories; (c) measures of community and community orientation; and (d) indicators of cultural values and tastes in food, music, and literature. Examples of the sorts of research questions this data may address include the following:

1. How have the lifetime migration histories of individuals changed over time? A number of large-scale data collection efforts look at whether individuals have moved during the past 5 or 10 years, and still other studies look at lifetime mobility histories for particular cohorts. Survey2000, however, allows researchers to study changes in lifetime mobility patterns across cohorts.
2. Where do individuals in today's society find a sense of community? Geographic communities, extended family, voluntary associations, workplace relations, and 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 United States 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 on the Internet, an area marked by great potential but also little experience (Blank, 1997; Fisher, Margolis, & Resnick, 1996; Schaefer & Dillman, 1998). From this experiment, we learn the following:

1. 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, to what extent can broad-based promotion and outreach efforts extend survey coverage to the general population?
2. 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.

SURVEY CONTENT

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. The project's academic collaborators developed the specifics of the survey instrument.

As noted above, Survey2000 consists of three main instruments: (a) the Canadian and U.S. Adult Survey, (b) the Youth Survey, and (c) the International Respondent Survey. On 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 reas-

signed based on this information. The major sections and key questions of each instrument are discussed below.

Canadian and U.S. Adult Survey

The Canadian and U.S. 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 U.S. 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, providing a greater range of response categories or an open-ended self-identification question was discussed. However, it was concluded 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.³ Respondents are also queried as to the frequency with which they engage in specific Internet activities (e.g., e-mail, purchasing products, use of listservs).

The next block of questions details each respondent's individual mobility history. Respondents are asked if they have ever lived outside the U.S. and Canada, how long they have lived at their current address, and whether they have always lived within 30 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 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 and territories associated with each. Respondents who were born in the U.S. or Canada are then sent to a further screen that asks the respondent to choose the state or province of their birth. Following up on this choice, the respondent is asked to select the closest location from a list of 10 to 15 cities in that state or province. Subsequently, respondents are asked a similar sequence of questions regarding their residence at ages 7, 14, 21, 28, 35, 42, 50, 60, 70, 80, and 90. 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.

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 e-mail) they have social contact with (a) relatives who live less than 30 miles away, (b) friends who live less than 30 miles away, (c) relatives who live more than 30 miles away, and (d) 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 reg-

ularly 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., signing petitions, voting, and attending a community or town meeting). This section then closes with a series of Likert scale items related to community (7-point scale, ranging from *strongly agree* to *strongly disagree*). These items⁴ 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) to 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, which are 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 music, literature, food, and views on the world. The literature and food modules are set up quite similarly. In each of these two modules, a customized list of items—authors in the case of literature and dishes in the case of food—are generated for each respondent based on their mobility history. These lists include 28 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 Real Audio sound clips that represent particular subgenres. The views-on-the-world topical module consists of 18 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.⁵

Although respondents initially receive 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 reward, 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 U.S. 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 U.S. and Canadian respondents. However, early in the project development process, it was decided that the project lacked the resources necessary to develop parallel instruments for all international respondents. Moreover, the orga-

nizational apparatus of the NGS, which is seen as a direct means to counter the bias inherent in a web survey, loses much of its efficacy outside North America.

Nonetheless, Survey2000 is a World Wide Web survey: respondents from 178 different countries and territories completed surveys. Anticipating this range of respondents, Survey2000 did not ignore respondents from outside North America or force them to respond to an instrument that showed little sensitivity to a significant subset of respondents. The decision was made to turn a potential limitation of Survey2000 into an advantage. Survey2000 offers a unique opportunity to empirically test the claim that American cultural imperialism has caused the emergence of a global *McCulture* at the expense of regional and national cultural diversity. Survey2000 oversampled 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 about 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 do 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 United States or any country other 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 U.S. and Canadian instrument, which customizes the literature and food modules according to each respondent's personal geographic mobility history.⁶ 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.

Youth Survey

All respondents under the age of 16, regardless of nationality, are directed to the youth survey. The survey content differs substantially depending on whether the respondent is between the ages of 13 and 15 or 12 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 questions for youth respondents focus on household composition and activities undertaken with parents and guardians. Children ages 13 through 15 are also asked about 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 ages 13 through 15. 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 social change in the next century.

THE WEB-BASED SURVEY DESIGN

As survey research has become increasingly sophisticated, social scientists 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.⁷ Survey research requires social interaction, which in turn, is sensitive to the technology on which the process rests. Different survey research techniques, including 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 & 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 & Pitkow, 1996).⁸

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.

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 about whether specific relatives live in their household. The greatest efficiency gains are realized in the geographic mobility section.

In the case of food and literature questions, using a web-based script supports flexibility that would never be possible in a self-administered or interviewer-assisted 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.

Design Elements

The design elements of Survey2000 are an essential feature of the project. Most fundamentally, the project's affiliation with NGS 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 2-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. Banner advertisements, widely used on commercial web sites including the NGS site, are not included on the Survey2000 pages so respondents would not think that the survey was a market research tool. Furthermore, the design goal was 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 he or she lived at birth or in a given year, the sidebar text reminds the respondent of the year in question and lists three events that took place in that year.⁹ These facts are designed as prompts but also to make the survey process 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, check boxes and radio buttons are used extensively to minimize the respondents' keyboard input.

Pretest 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 randomly 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, which included the option to continue the survey and respond to the three remaining topical modules. This approach might have had costs of its own with respect to possibilities for analysis across cultural domains; however, more than 70% of U.S. adults ($N = 23,384$) voluntarily continued after completing the base module ($N = 32,688$). Thus, with this group of respondents, 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

From a survey research perspective, the nonrandom nature of a web survey sample raises serious questions; however, this issue 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: Survey2000 relies on (a) items from existing surveys conducted with traditional sampling and survey methods (e.g., the GSS) to provide external benchmarks to assess the nature of the survey bias and construct the necessary weights and (b) extensive use of NGS public relations and community outreach resources to extend survey coverage. To begin with, however, several characteristics of the Survey2000 sample are noteworthy.

Tables 1 and 2 provide an overview of the entire Survey2000 sample and their form of participation. More than 80,000 surveys were initiated, and a bit more than 50,000 were completed. Adults living in or citizens of the U.S. or Canada initiated the most surveys ($N = 45,951$) and completed 37,091 surveys. Adults from other parts of the world comprised about one fifth of the initiated surveys. Youth surveys were completed by 9,785 children in the United States, 970 Canadian children, and 1,635 international youth. The overall survey completion rate was greater than 70% for all adults and almost 60% for children.

Approximately half of the initiated surveys ($N = 80,015$) were from U.S. adult respondents ($N = 40,612$). Combined with an 80.5% completion rate, this yields a sample of 32,688 complete U.S. adult surveys. For many of the questions these data address, partial responses are of interest. Thus, a sample size of more than 32,688 respondents exists for demographic

TABLE 1
Survey2000 Participation and Completion by Survey Form

	<i>Number of Surveys Initiated</i>	<i>Number of Surveys Completed^a</i>	<i>Percentage Completed</i>
All adults (age 16 or older) ^b	65,676	47,176	71.8
All youth (age 5 though 15)	14,339	7,761	54.1
U.S. adults (age 16 or older)	40,612	32,688	80.5
U.S. youth (age 5 though 15)	9,785	6,246	63.8
Canadian adults (age 16 or older)	5,339	4,403	82.5
Canadian youth (age 5 though 15)	970	633	65.3
Other international adults (age 16 or older)	13,613 ^c	10,085	74.1
Other international youth (age 5 though 15)	1,635	882	53.9
Total	80,015	54,937	100.0

a. A survey is considered completed if the respondent completed the base module to the point at which the first of four cultural modules began. See Table 2 for a more detailed breakdown of completion rates for U.S. adults.

b. Includes 6,112 adult and 1,949 youth respondents who dropped out prior to reporting residence.

c. Total includes 961 (7.1% of total) respondents who are U.S. or Canadian citizens living abroad.

TABLE 2
Participation and Completion Rates for U.S. Adults

<i>Level of Participation</i>	<i>Number of U.S. Adults^a</i>	<i>Percentage</i>
Volunteered for survey	40,612	100.0
Volunteered but did not complete the base module	7,924	19.5
Completed base section (which included a culture module) and did not continue	9,117	22.4
Completed base section and volunteered to continue but only partially completed the extra culture modules	3,222	7.9
Completed base and all four culture modules	20,349	50.1

a. The number of respondents who volunteered for the survey (40,612) minus the number who did not complete the base module (7,924) yields 32,688. This is the number of U.S. adults who completed the survey as noted in Table 1.

and social capital measures, as well as the Internet access and use items, that begin 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 3, which compares the Survey2000 sample with recent GSS surveys and U.S. 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 need to be statistically adjusted to adequately represent the U.S. population. Second, the Survey2000 sample provides some insight into the magnitude of the difference between the general U.S. population and its Internet population.

Recent findings indicate that the gender gap in Internet use for U.S. adults has essentially disappeared, and trends toward equal access by education, income, and race have also been noted (Clemente, 1998; Glasner, 1999; Katz, 1997). Particularly concerning gender, our results corroborate these findings. This comparison shows, for example, that whereas just greater than half (50.7%) of the Survey2000 sample is male, female respondents constitute

TABLE 3
Demographics of the Survey2000 Adults (age > 18) sample compared to the
1996 and 1993 General Social Surveys and Census Bureau Statistics^a

	Survey2000 ^b		1996 General Social Survey		1993 General Social Survey		1997/1998 Census Bureau ^c	
	%	N	%	N	%	N	N (in % thousands)	
Gender								
Female	48.9	15,147	55.7	1,614	57.3	918	51.9	100,954
Male	51.1	15,801	44.3	1,283	42.7	683	48.1	93,474
Median age in years	38		44		43		40-44	
Race								
Black	1.4	428	13.9	402	11.2	179	11.6	22,590
White	94.5	29,004	80.9	2,344	83.9	1,343	84.0	163,368
Other	4.1	1,268	5.2	151	4.9	79	4.4	8,472
Education								
Less than high school degree	0.9	292	15.2	441	18.1	289	17.9	35,246
High school degree	31.9	9,882	54.1	1,567	52.5	840	52.9	104,334
Associate's degree	7.8	2,421	6.7	194	6.2	99	7.1	13,996
Bachelor's degree	34.1	10,569	16.3	471	15.8	253	15.2	30,087
Graduate degree	25.2	7,785	7.7	224	7.4	118	7.0	13,750

SOURCE: Data sources for 1997/1998 Census Bureau: Gender data for the year 1997 is from U.S. Census Bureau (1998, No. 15, p. 16). Race data for the year 1997 is from U.S. Census Bureau (1998, No. 22, p. 22). Education data for the year 1998 is from U.S. Census Bureau, Population Division (1998, pp. 2-6).

a. Sample is restricted to age 19 or older to facilitate GSS comparison for all data sources.

b. There were also 713 respondents who did not provide information on race.

c. Education numbers include those 18 years and older

the majority in the 1996 (55.7%) and 1993 (57.2%) GSS samples.¹⁰ Furthermore, 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 underrepresented on the Internet: 92.5% of the respondents are White.¹¹ Only 1.5% of the U.S. adult surveys are from African Americans. In subsequent efforts to generalize from Survey2000 to the broader U.S. population, weights will need to be developed to make the necessary statistical adjustments. But it should also be pointed out that the large sample size should make this possible. Although only 1.5% of the respondents are African American, this amounts to 538 surveys. The 1993 GSS only has 179 African Americans, whereas the GSS African American 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. In this regard, the actual proportions across categories are not as important as the number of respondents within each cell.

Finally, Table 3 indicates a large difference between the educational makeup of the Survey2000 sample and the U.S. population at large. Only 0.9% 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 U.S. 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 postsecondary degree (31.9%) is considerably lower than that found in the

1996 GSS (54.1%) and the 1993 GSS (52.5%). Correspondingly, respondents with postsecondary degrees are overrepresented 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 U.S. population at large. In and of themselves, these numbers reveal a great deal about the educational background of the Internet community and the educational aspects of the current digital divide.

SAMPLING ISSUES: RANDOMNESS AND REPRESENTATION

The potential for sample bias in data collected from the Internet represented the most serious methodological problem 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 of more than 2 million but still came to the wrong conclusion.¹² 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, whereas 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 oversampled 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 and even the size and boundaries of population membership are unknown. Nonetheless, this does not mean that the survey cannot yield representative social science data. To begin with, although 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 NLSY79. These studies are based on traditional sampling and data collection methods. Results based on these other items may serve as external benchmarks as well.

Based on such 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. Because 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 less than a dozen members, whereas 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 or 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, in 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, p. 43).¹³ For certain subgroups—for example, the educated, African American females, those older than age 60, those 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 NGS's participation in this project is so important. The society used other NGS media—the magazine, television, and school-based educational activities¹⁴—to encourage participation from a wider range of people than those normally found on the web.

Because much of the Survey2000 sample was generated by NGS, we acknowledge a possible sample bias in that respondents will probably have many of the attributes of typical visitors to the NGS web site. However, publicity was generated over listservs and through arti-

cles in several magazines and newspapers. For example, during a 2-day period in which *HotWired* magazine provided a direct link to the survey, some 2,600 surveys were initiated, although, 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 generated publicity and broadened the sample diversity.

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, NGS 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.

Clearly, the Survey2000 sample selection 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. Because we are interested in the ability of Survey2000 data to represent both the Internet population and the general U.S. population, we compare our sample with kinds of samples of these two populations.

To assess the quality of our sample with regard to the Internet population, 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 state.¹⁵

A second, more controversial issue regarding the Survey2000 sample concerns the extent to which it represents the population off-line. The results presented in Table 3 clearly indicate that particular demographic groups are strongly overrepresented in the sample and others notably underrepresented. The key question then becomes whether demographic 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, do the Survey2000 and GSS population of single, White males between the ages of 19 and 40 with at least a bachelor's degree have similar musical tastes?¹⁶ The comparison is not perfect because the GSS data was collected in 1993, 5 years earlier than Survey2000. Thus, shifting musical tastes among members of this subgroup may 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 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 20 genres of music, and 16 of these closely correspond to categories used in the 1993 GSS. Table 4 presents comparisons using seven illustrative music genres for the GSS and Survey2000 subsamples of White males between the ages of 19 and 40 with a bachelor's degree. Considering these respondents' reactions to the big band/swing genre, a clear gap is evident: Of the GSS respondents 7.9% "like it very much," and another 48.3% "like it," whereas 24.9% of the Survey2000

TABLE 4
Musical Tastes Among White Respondents Between the Ages of 19 and 40
With a Bachelor's Degree (in percentages)

	<i>General Social Survey</i>	<i>Survey2000</i>
Big band		
Like very much	7.8	24.9
Like it	48.3	42.8
Mixed feelings	24.1	22.6
Dislike it	11.2	6.4
Dislike very much	2.6	1.7
Don't know much about it	6.0	1.6
Country and western		
Like very much	12.1	6.7
Like it	25.9	17.1
Mixed feelings	39.7	31.5
Dislike it	16.4	22.3
Dislike very much	5.2	21.8
Don't know much about it	0.9	0.6
Blues or rhythm and blues (R & B)		
Like very much	13.8	18.4
Like it	44.0	41.7
Mixed feelings	27.6	27.8
Dislike it	9.5	7.5
Dislike very much	4.3	2.0
Don't know much about it	0.9	2.6
Jazz		
Like very much	19.8	25.8
Like it	37.1	38.2
Mixed feelings	29.3	25.7
Dislike it	8.6	6.7
Dislike very much	4.3	2.3
Don't know much about it	0.9	1.2
Classical/symphony and chamber music		
Like very much	25.9	41.1
Like it	32.8	40.9
Mixed feelings	31.9	14.3
Dislike it	5.2	1.9
Dislike very much	4.3	0.7
Don't know much about it	1.2	
Latin		
Like very much	2.6	10.7
Like it	25.9	38.6
Mixed feelings	25.9	30.9
Dislike it	23.3	10.7
Dislike very much	8.6	2.3
Don't know much about it	13.8	6.8
Contemporary rock		
Like very much	22.4	24.5
Like it	56.0	44.4
Mixed feelings	14.7	24.2
Dislike it	3.4	4.6
Dislike very much	3.4	1.4
Don't know much about it	0.0	0.8
Total N	116	3,003

respondents "like it very much" and 42.8% "like it." However, big band/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 big band/swing music among educated, young, White males.

Considering blues and rhythm and blues (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.¹⁷ Significantly larger differences between the two samples are found comparing feelings regarding classical music. Among GSS respondents, 58.7% reacted favorably to classical music, whereas 82.0% 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-and-western genres than did 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, whereas weighting down those respondents with dissimilar preferences should capture some of the unobserved selection differences between the two samples. Although this approach is far from perfect, it does afford one means to further generalize from Survey2000 to 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 voluntaristic sampling approach along with external benchmarks and an aggressive outreach effort to diversify the sample was employed. In our view, the unweighted Survey2000 data is likely to provide fascinating insights into the demography and sociology of the broad mainstream of web users. We also believe that the Survey2000 data may also be used as a data source for the U.S. population at large; however, for these purposes, the selective nature of the sample needs to be taken into account. Furthermore, beyond substantive findings from Survey2000, the project offers important methodological lessons. There has been little empirical research in large-scale web survey administration. These data provide a solid foundation on 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 foreshadows an era when clickable maps to record respondent geographic mobility and widespread use of other multimedia 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 is especially important because 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.

NOTES

1. The Survey2000 World Wide Web site can be visited at <http://survey2000.nationalgeographic.com/index.html>.

2. The Survey2000 data, codebook, and other supporting documentation may be obtained at <http://business.clemson.edu/socio/s2kdata211.htm>.

3. 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 (although 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.

4. These items include the following: 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.

5. 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.

6. 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.

7. The significance of this relationship is noted quite clearly by Alfred Schutz (1971) 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."

8. Another interesting Internet survey project is hosted by the Graphics, Visualization, and Usability Center (GVU) of the Georgia Institute of Technology, which has been conducting web-based surveys for 5 years. The GVU's 10th survey offered cash incentives to respondents, advertised corporate sponsorship, and collected more than 5,000 responses. Whereas the GVU 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.

Whereas GVU 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.

9. 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 2 months apart. Film director Stanley Kubrick releases *2001: A Space Odyssey*. U.S. first-class postage: 6 cents.

10. A long-standing issue with the General Social Survey (GSS) and other probability samples has been the overrepresentation of females (Smith, 1997).

11. 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.

12. An oft-overlooked fact is that in the previous four presidential elections, Literary Digest Polls correctly predicted the winner. Beyond the 1936 truism that a large sample does not guarantee accurate results, it ought to be emphasized that a nonrandom sample does not amount to a recipe for invalid results.

13. Thinking solely about important demographic attributes—gender, age (four categories), marital status (three categories) race/ethnicity (three categories), educational achievement (three categories), employment status (two categories) and urban-suburban/rural residence (two 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 analysis. This is often the case, and many “nationally representative samples,” including GSS, are much smaller.

14. For example, the National Geographic Society’s Geography Education Program included an overview of Survey2000 as part of its annual summer geography program to more than 200 geography teachers the month before the survey went online.

15. 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.

16. The subgroup of single, White males between the ages of 19 and 40 with at least a bachelor’s degree represents the subgroup where the Survey2000’s large sample size offers the best possibilities for weighting and adjustment due to the large number of respondents within this demographic subgroup.

17. Survey2000 respondents are somewhat more likely to say they “like it very much” than are 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 & Sproull, 1986).

REFERENCES

- Blank, G. (1997). The road ahead: Observations on the role of the Internet. *Social Science Computer Review*, 15(2), 190-195.
- Bratton, G. R., & Newsted, P. R. (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, P. (1998). *The state of the Net: The new frontier*. New York: McGraw Hill.
- Fink, A. (1995). *How to sample in surveys*. Thousand Oaks, CA: Sage.
- Fisher, B., Margolis, M., & Resnick, D. (1996). Breaking ground on the virtual frontier: Surveying civic life on the Internet. *The American Sociologist*, 27(1), 11-29.
- Glasner, J. (1999). Gender gap? What gender gap? *Wired Magazine*. San Francisco, CA: The Condé Nast.
- Katz, J. (1997, March-April). The social side of information networking. *Society*, 34, 9-12.
- Kehoe, C. M., & Pitkow, J. (1996). Surveying the territory: GVU’s five WWW user surveys. *World Wide Web Journal*, 1(3), 77-84.
- Kiesler, S., & Sproull, L. S. (1986). Response effects in the electronic survey. *Public Opinion Quarterly*, 50, 402-413.
- Schaefer, D. R., & Dillman, D. A. (1998). Development of a standard E-Mail methodology. *Public Opinion Quarterly*, 62, 378-397.
- Schutz, A. (1971). *Collected papers* (M. Natanson, Ed.) (Vol. 1). The Hague, the Netherlands: M. Nijhoff.
- Smith, C. B. (1997). Casting the Net: Surveying and Internet population. *Journal of Computer-Mediated Communication*, 3(1). [Online] Available: <http://www.ascusc.org/jcmc/vol3/issue1/smith.html>
- U.S. Census Bureau. (1998, September 30). *The official statistics: Statistical abstract of the United States, 1998*. Washington, DC: Author.

U.S. Census Bureau, Population Division (1998, October). *Current population report. Education attainment in the United States: Detailed tables March 1998*. Washington, DC: Author.

James C. Witte is an assistant professor of sociology at Clemson University. He holds a doctorate in sociology from Harvard University and specializes in the sociology of work, demography, and research methodologies. He is the principle investigator on the Survey2000 project and can be contacted by e-mail at jwitte@clemson.edu.

Lisa M. Amoroso is a doctoral candidate in sociology and organization behavior at Northwestern University. Her research interests include the sociology of organizations, gender, and comparative state politics. Her dissertation explores the structural basis for gender inequality in a comparative framework, with special attention to methodological issues surrounding theory advancement. She can be contacted by e-mail at amoroso@nwu.edu.

Philip E. N. Howard is a doctoral candidate in sociology at Northwestern University. He specializes in the political sociology of development in poor countries, research methods, and the sociology of communication. His dissertation is about the effect of hypermedia technologies on deliberative democracy in the United States. He can be contacted by e-mail at p-howard@nwu.edu.