An Assessment of the Generalizability of Internet Surveys

Samuel J. Best, Brian Krueger, Clark Hubbard and Andrew Smith

Social Science Computer Review 2001 19: 131
DOI: 10.1177/089443930101900201

An Assessment of the Generalizability of Internet Surveys

SAMUEL J. BEST
University of Connecticut

BRIAN KRUEGER
University of Notre Dame

CLARK HUBBARD
ANDREW SMITH
University of New Hampshire

The Internet has become an increasingly popular form of data collection because it permits complex questionnaires to be administered more quickly, flexibly, and inexpensively than conventional survey methods. However, the Internet is restricted to individuals with access to computer networks. Thus, causal inferences to the general population from analyses of Internet samples necessarily rest on two untested assumptions: (a) that the decision-making processes of Internet users are similar to those used by the general population, and (b) that representative samples of Internet users can be drawn. The authors provide mixed support for these assumptions. They find that current Internet sampling techniques only permit the generation of diverse, not representative, samples. However, comparing samples drawn simultaneously using the Internet and probabilistic telephone methods, the authors demonstrate that the psychological mechanisms underlying common political decisions do not differ between Internet users and the population. They discuss the implications of these findings for future survey research.

Keywords: Internet, sampling, coverage error, inference

The Internet offers unprecedented opportunities for data collection. It provides access to millions of potential research participants, including populations with special characteristics (Binik, Mah, & Kiesler, 1999; Schmidt, 1997). It permits complex instruments capable of experimentally manipulating stimuli, accommodates audio and video transmissions, and facilitates live interaction between participants (Hewson, Laurent, & Vogel, 1996; Smith & Leigh, 1997). And it can be employed quickly, conveniently, and inexpensively by eliminating the need for interviewers or synchronous interaction (Mavis & Brocato, 1998; Schaefer & Dillman, 1998; Sproull, 1986).

However, the Internet suffers from one formidable disadvantage. There is not universal access to the medium. As of December 2000, 43% of the population were not connected to the Internet (Nielson//NetRatings, 2000). Moreover, the backgrounds of Internet users differ significantly from nonusers, exhibiting higher education and socioeconomic status than their counterparts (National Telecommunications & Information Administration, 2000). As a

Social Science Computer Review, Vol. 19 No. 2, Summer 2001 131-145

result, samples of Internet users are vulnerable to systematically ignoring certain attributes of nonusers and generating misleading conclusions about the general population.

The threats from coverage error, though, have not discouraged social scientists from using electronic communication for data collection. Researchers have analyzed the decision-making processes of Internet users on topics ranging from politics (Bimber, 1998) to prettiness (Krantz, Ballard, & Scher, 1997). Results from such analyses have frequently been used to make inferences about the psychological mechanisms underlying much broader populations—populations often comprising individuals without Internet access.

The ability of researchers to make such generalizations rests on two untested assumptions implicit in their analyses. First, they assume that the decision-making processes of Internet users are no different from those in the population at large. Second, they assume that representative samples of Internet users can be drawn. In the absence of either of these assumptions, cross-sections of Internet users would be reduced to nothing more than convenience samples from which statistical inferences could not be made about any tendency of the general population.

This article assesses the generalizability of Internet surveys by testing the validity of these assumptions. We find evidence that only diverse, not representative, samples of Internet users can be generated. However, comparing samples drawn simultaneously using the Internet and probabilistic telephone methods, we find that the psychological mechanisms underlying common political decisions do not differ between Internet users and the population. Thus, we conclude that the use of Internet samples should be limited to those circumstances in which some demonstrable evidence exists that the decision-making hypotheses being tested are uniformly applicable to the entire population.

RECOVERING FROM COVERAGE ERROR

Enthusiasm for the Internet has grown exponentially in recent years. From 1995 to 2000, the number of Americans with access to the Internet increased from 22 million to 158 million (Nielson//Netratings, 2000). Despite its growing popularity, substantial numbers of people still do not have access to the Internet. Demographic information about Internet users suggests that their backgrounds differ significantly from those without connections to the Internet. Internet users are more likely to be young, White, married, higher educated, and wealthier compared with the population at large (National Telecommunications & Information Administration, 2000). Such coverage error raises serious questions about our ability to infer characteristics of the general population from analyses of Internet samples.

Confidence in the representativeness of survey marginals requires that every unit in the target population possesses some chance of being selected so that the statistical likelihood of drawing each population unit can be computed (Babbie, 1990). Such probability sampling has the advantage that when samples of a given size are drawn repeatedly, corresponding sample estimates will form distributions from which the true population parameters can be derived. By exploiting the statistical properties of these distributions, researchers can calculate with a degree of certainty how much sample estimates deviate from the true population parameters (Judge, Hill, Griffiths, Lütkepohl, & Lee, 1988). In other words, probability sampling enables researchers to generate confidence intervals within which population parameters possess some statistical likelihood of being found.

In the absence of probability sampling, researchers cannot judge the probable accuracy of univariate estimates. Nonprobability samples may systematically ignore certain values of particular measures, meaning that the probability of their occurrence in the population cannot be computed. As a result, researchers cannot calculate the statistical likelihood that con-

fidence intervals will contain the true population parameters (Judge et al., 1988). Without such assurances, generalizations from nonprobability samples are no different from educated guesses about the nature of the population.

The purpose of collecting data, though, is not always to predict univariate parameters. Many of our questions focus on testing causal hypotheses through our analysis of multivariate relationships. Researchers can draw statistical inferences about the causal mechanisms underlying the general population from samples of Internet users if they are willing to make two assumptions. First, they must assume that the decision-making processes used to generate a given attitude or belief are the same for Internet users as nonusers. In other words, the psychological mechanisms that produce certain beliefs or attitudes are not a function of Internet use. In practice, this means that the experience of using the Internet does not generate or prime beliefs or attitudes that would directly influence the dependent variable under investigation or indirectly mediate or moderate how other factors influence this dependent variable.

To illustrate, consider that individuals in the general population form approval ratings of the sitting president from their party identification and their perception of the president's handling of the economy. To rely on a sample of Internet users, then, it must be assumed that this medium does not alter this psychological mechanism. Therefore, within an Internet sample, party identification and economic performance should influence presidential approval in the same manner. Different factors should not influence presidential approval nor should they mediate or moderate the manner in which party identification and economic performance influence presidential approval. If they do, researchers could not extrapolate information about the general population from an Internet sample.

The second assumption necessarily made by researchers employing Internet samples for causal hypothesis testing is that a representative sample of Internet users can be drawn. This means that a sampling frame can be constructed that gives every Internet user some chance of being selected. Because Internet addresses are not compiled in a single registry nor do they adhere to a standard format, we cannot generate a simple random sample. This poses a serious challenge for researchers to identify a sampling unit that yields a statistical probability of selecting each Internet user.

Three different sampling units are frequently used to construct sampling frames of the Internet population: listed e-mail addresses, electronic subscription groups, and heavily trafficked Web sites. E-mail addresses are frequently compiled in online directories that function as the near-electronic equivalent of a telephone book. They range from institutional rosters to public service directories. Researchers select every address, or at least a random subset, from one or more of these directories and treat them as analogous to probabilistic samples of Internet users (Gotcher & Kanervo, 1997; Stern & Faber, 1997). Thus, concerns over representativeness are seemingly rendered irrelevant by replacing the population of Internet users with a complete list of e-mail addresses.

Alternatively, electronic subscription groups are used as sampling units of Internet users (Bonchek, Hurwitz, & Mallery, 1996; Fisher, Margolis, & Resnick, 1996; Swoboda, Mühlberger, Weitkunat, & Schneeweib, 1997; Wherrett, 1999). Whereas there is no exhaustive directory of e-mail addresses, there are fairly comprehensive inventories of electronic subscription groups. Organizations compile files of subscription groups such as Usenet newsgroups or Listserv mailing lists. Rather than identifying individual users, researchers randomly sample groups from these inventories and then send the survey to subscribers of the selected groups in much the same way that cluster sampling is applied to the general population. Thus, all groups, rather than subscribers, are believed to have the same chance of being selected.

Finally, heavily trafficked Web sites are used to construct sampling frames of the Internet population (Bimber, 1998; Georgia Institute of Technology, Graphic, Visualization, & Usability Center, 1998; Johnson & Kaye, 1998; Wherrett, 1999). Employing these sampling units works similarly to using electronic subscription groups. Researchers generally select Web sites from among those that have generated a threshold number of downloads per day or per month. After these Web sites have been identified, advertisements are posted that direct interested visitors to a separate Web site that contains the survey.

Regardless of the approach taken, the validity of the assumptions underlying the use of Internet samples for causal hypothesis testing has not been tested. Can a representative sample of Internet users be drawn using the sampling procedures currently available? Do Internet users rely on different psychological mechanisms to process information than those used in the public at large? In the absence of these assumptions, can analyses of Internet samples be generalized? This article pursues such questions.

ASSESSING THE REPRESENTATIVENESS OF INTERNET SURVEYS

Currently, no support exists for the assumption that a representative sample of Internet users can be drawn. Regardless of the sampling unit employed, a sampling frame cannot be designed that ensures that each Internet user possesses some chance, even if not an equal one, of being selected. It does not matter which selection criteria—random, stratified random, or cluster drawing—researchers use, the sampling frame will not match the target population of Internet users exactly.

Sampling frames using listed e-mail addresses exclude numerous Internet users who do not appear in online directories, regardless of the number of addresses or which organization manages the list. Online e-mail directories frequently rely on volunteer, not automatic, registration. As a result, e-mail users unwilling to participate will not appear in such listings. Moreover, online directories do not list the approximately 20% of Internet users who do not have e-mail accounts (National Telecommunications & Information Administration, 2000). Even if an overwhelming number of individuals had active accounts appearing in a directory, there is no comprehensive listing of e-mail directories. With countless directories in existence, researchers could not even judge the likelihood that they sampled a subset of such directories.

Sampling frames relying on electronic subscriber groups do not fare any better. Although comprehensive inventories of subscriber groups may exist, although even this is open to question, numerous Internet users do not take part in newsgroups or mailing lists and have no opportunity to be selected. Moreover, the number of subscriber groups changes daily as new groups are created and old ones removed. Thus, the sampling frame not only deviates from the target population but the degree to which it deviates changes regularly, making the problem nearly impossible to overcome.

Sampling from Web sites, regardless of the amount of traffic, suffers from many of the same problems as sampling from electronic mailing lists. At any given interval, countless Internet users will not visit survey-linked Web sites either by intent or coincidence, a problem that is only magnified when eligible Web sites require a threshold amount of traffic. Moreover, both the number of Web sites and the users trafficking each Web site are constantly changing over time, meaning sampling frames of Web sites will also differ from one moment to the next.

Because sampling frames based on each type of sampling unit deviate from the target population, it is impossible to ensure that Internet users excluded from the sampling frame do

not differ systematically on certain hypothetical measures from those included in the sampling frame. It is conceivable that accessible Internet users rely on psychological mechanisms different from inaccessible Internet users.[1] The causes of an Internet user's inaccessibility or the beliefs and attitudes generated or primed by the function of being accessible may alter the means by which these groups process information. Without being able to contact inaccessible Internet users, there are no means of rectifying, or for that matter even identifying, such possibilities.

However, if we are willing to assume that the psychological mechanisms are the same for accessible and inaccessible Internet users as well as for Internet users (accessible and inaccessible) and nonusers, then we would not need representative samples of the Internet population in order to extrapolate to the general population. A sample simply containing a dispersion of values for all of the variables in the hypothesized model would be sufficient. The rationale is that the observed differences between the marginals of the univariate distributions across sampling modes reflect shifts in the position of the variables along their respective axes and do not change the correlations between the variables. (De Leeuw, Mellenbergh, & Hox, 1996; Krosnick & Alwin, 1988). As long as some variance exists for all the variables in the hypothesized mechanism and the missing variance does not correlate differently than the measured variance, then a diverse rather than a representative sample can be used to infer relationships within the population.

Such reasoning is similar to that which underlies the use of undergraduates in experimental psychology (Lang, 1996; Sears, 1986). University undergraduates are frequently used to evaluate the effects of certain stimuli on decision making. Although they randomly assign participants to various experimental groups, researchers do not randomly select undergraduates but rely on convenience samples of available students. In doing so, they must assume that these available undergraduates process information similarly to unavailable undergraduates, as well as like others in the general population; otherwise, their conclusions about the reactions of undergraduates to stimuli cannot be generalized to the population (Anderson, Lindsay, & Bushman, 1999).

Internet researchers have already begun relying on this rationale (Buchanan & Smith, 1999; Davis, 1999; Krantz et al., 1997; Smith & Leigh, 1997; Witte, Amoroso, & Howard, 2000). Samples are typically drawn from chance visitors to the Web site where the survey is posted. Researchers collect responses until participants' demographics closely resemble data from existing probabilistic telephone samples. Analyses of these convenience samples of Internet users are then generalized to the population.

Nevertheless, if Internet users do not rely on similar psychological mechanisms as nonusers to generate beliefs or attitudes, then even using extremely diverse samples of Internet users would not be sufficient to infer conclusions from them to the general population. In the next section, we empirically evaluate this assumption. We test the similarity in decision making between Internet users and the general population by comparing the results of a political survey conducted using both the Internet and probabilistic telephone methods.

DATA COLLECTION

In the spring of 1999, we simultaneously drew a convenient Internet sample and a probabilistic telephone sample of registered voters residing in New Hampshire. We then compared the psychological mechanisms underlying two common political decisions. In an effort to make a fair and meaningful comparison of the effects of using each sampling mode, we tested simple, well-understood hypotheses using conventional interviewing practices established for each form of data collection.

Questionnaire Design

The questionnaire sought to understand the voting intentions of registered New Hampshire voters in the upcoming presidential election. After a brief introduction, we commenced the survey with questions assessing respondents' likely participation in the 2000 presidential election. Next, respondents evaluated the performance of President Clinton and indicated their presidential preferences in a series of mock elections. We then gauged the salience of and support for various government policies. Finally, we asked respondents a number of questions tapping their political orientations and demographic characteristics.

With few exceptions, the survey questions were identically worded and ordered on the e-mail and telephone versions. The design and layout of the electronic version of the survey aspired to simplicity and clarity. The questionnaire appeared as black, 14-point, Times New Roman print on a white background. It contained no color or graphics except for a small university logo in the upper left-hand corner of the survey. Text lines were approximately three quarters of the width of a standard computer monitor to avoid wraparound on most screen configurations. Questions contained drop-down menus showing relevant choices. After completing the questions, respondents clicked the Submit button at the end of the questionnaire to return the survey. On transmitting the survey, respondents received an acknowledgment that the data had been sent.

Sampling

We generated a probabilistic sample of telephone numbers by using an unrestricted sampling process within explicit sampling intervals. Initially, working area code/exchange/bank combinations in New Hampshire were identified. Next, numbers were generated by randomly selecting prefixes from working combinations along with two-digit suffixes. Finally, these numbers were screened to remove nonresidential businesses and nonworking numbers by simultaneously comparing them to business listings in the New Hampshire Yellow Pages and residential listings in the New Hampshire white pages.[2]

Recognizing that we could not generate a representative sample of Internet users, we sought to produce a diverse sample with a broad dispersion of responses for each of the questions in our survey. We constructed our Internet sample by extracting e-mail addresses originating in New Hampshire households from several prominent public service directories on the World Wide Web (WWW). We made no effort to truncate the e-mail addresses by random selection or other means, but used all of the addresses available.

Survey Administration

Telephone numbers were dialed daily between 10:00 a.m. and 9:00 p.m. from April 24 to May 2, 1999 by trained interviewers from 24 centrally supervised phone banks.[3] Interviewers attempted up to 10 callbacks to contact potential respondents, averaging 2.4 callbacks per number. Answering machines were treated as "no answers" and called back on the regular no-answer rotation a minimum of 3 hours later. Senior interviewers contacted refusals on one additional occasion for conversion, which proved to be successful roughly a quarter of the time. After securing cooperation, interviewers used Computer-Assisted Telephone Interviewing (CATI) systems to administer questions and record responses.

E-mail addresses were transmitted a prenotification 2 days prior to the dissemination of the survey. The prenotification described our intentions, provided the name and e-mail address of a contact for further information, and contained a hyperlink connecting to a

WWW page of responses to frequently asked questions. On April 24, 1999, we transmitted a cover letter to the e-mail sample. The cover letter encouraged recipients to participate in the study and included a hyperlink to the actual survey written in HyperText Markup Language (HTML), accessible by most current browsers. E-mail surveys were accepted through May 2, 1999. We ensured the confidentiality of the participants by replacing all identifying information with an identification number after eliminating multiple submissions but before merging it with other responses in an aggregate computer file.

MODELS AND ANALYSIS METHOD

To investigate effects of sampling mode on preference formation, we examined two common political decisions: presidential approval and vote choice. For our objectives, the structure of the causal mechanisms was not of paramount importance. However, because we wanted to compare parameter estimates across the two survey modes, it was useful to test models with satisfactory overall fits. Therefore, we relied on simple variations of existing models for which prior research generated strong fits.

Modeling Presidential Approval

Our equation for presidential approval, derived from research by Gilens (1988) and Hurwitz and Peffley (1987), treated approval as a function of retrospective evaluations of the president's economic and foreign policy performance, positions on contemporary policy issues, and long-standing political predispositions. We measured presidential approval by asking respondents to answer the question, "Generally speaking, do you approve or disapprove of the way President Clinton is handling his job as president?" using a 7-point response scale anchored by *strongly approve* and *strongly disapprove*. Retrospective performance evaluations were assessed by asking respondents to indicate on the same 7-point scale whether they approved or disapproved of the way President Clinton handled the economy and the way he handled foreign affairs. To assess policy preferences, we asked respondents their positions on two salient political issues: abortion and sending troops to Kosovo. The abortion question asked respondents to choose whether they believed abortion should be *legal in all circumstances*, *allowed only in cases of rape, incest, or to save the life of the mother*, or *not legal in any circumstances*. We gauged support for Kosovo troop intervention by assessing respondents' degree of support for "sending U.S. ground troops into Kosovo with troops from other NATO countries," using a 5-point response scale ranging from *strongly favor* to *strongly oppose*. We relied on trichotomous measures of partisanship and ideology to tap respondents' long-standing political dispositions. Finally, we included controls for respondents' age, education, income, gender, marital status, parental status, religious attendance, and newspaper readership.[4]

Modeling Vote Choice

Our two vote-choice equations, stemming from work by Abramowitz (1995) and Shanks and Miller (1990), modeled candidate preference as a function of presidential approval, policy positions, and demographics. In the first equation, the dependent variable was measured as the respondent's candidate preference in a hypothetical presidential election between Al Gore and George W. Bush, the front-runners for the Democratic and Republican parties at the time of the survey. The second equation matched Elizabeth Dole, the primary challenger of George W. Bush for the Republican nomination at the time, against Al Gore. We included

the same independent variables in both equations—Clinton approval, positions toward abortion and Kosovo involvement, partisan and ideological identification, and demographic characteristics—relying on the identical questions to assess them as were used in the presidential approval model.

Method

We used the same evaluative strategy for each of the decision-making processes.[5] To begin, logistic equations were fit to the dependent variables using combined data from both the e-mail and the telephone samples. Ordered logit was used to assess presidential approval, whereas binary logit was used to evaluate the vote choice models. These equations served as baselines from which comparisons of sampling modes were made.

Next, a multivariate test for sampling mode effects was employed in each of the models. Models were fit with all the additive terms from the baseline model, a dummy variable for the sampling mode (0 = Internet, 1 = telephone), and multiplicative terms interacting the sampling mode dummy with each variable in the baseline model. A simultaneous joint chi-square test of all the interaction parameters was used to determine if the overall relationships between the variables differed between the telephone and e-mail samples (see Aquilino, 1994). Similar to the chi-square test used to determine the significance of many maximum likelihood models, a significant test for the group of parameters would indicate broad inequality across the sampling modes.

Although the overall causal connections between variables may be statistically equivalent across sampling modes, some consistently unequal parameters may remain hidden when using a joint chi-square test. To guard against possible drowning effects from this omnibus test, we concluded our comparisons with the strictest test of equality in which each interaction term and mode dummy were separately entered into the baseline model. In other words, rather than testing the parameters as a whole by simultaneously including all the interaction terms into the model, we tested the equality of each parameter by individually entering each interaction term into the baseline model (Maddala, 1992). A significant interaction term would indicate unequal relationships between the e-mail and phone sample for that particular variable.

RESULTS

Internet Sample Diversity

Table 1 reports measures of dispersion for each of the hypothesized variables in the Internet and probabilistic telephone samples. It provides strong evidence that we generated a diverse sample of Internet users. For the categorical dummy variables, we find that at least a third of the Internet sample falls into each category. For the ordered variables, we find standard deviations similar in magnitude to those found in the telephone sample.

The Approval Model

Table 2, column 1, shows the results of the baseline model for presidential approval. Judgments of the president's economic and foreign policy performance, newspaper readership, party identification, and education all proved to be significant determinants of respondents' presidential evaluations. To determine whether these cumulative results differ from our

TABLE 1
Central Tendency and Dispersion of Variables in the Telephone and Internet Samples

	Central Tendency		*Dispersion*	
	Telephone	*Internet*	*Telephone*	*Internet*
Categorical				
Child	0 (60.7%)	0 (50.2%)	.954	.999
Married	1 (59.4%)	1 (68.2%)	.965	.868
Gender	1 (57.2%)	0 (65.3%)	.979	.906
Gore/Bush	1 (61.7%)	1 (59.1%)	.945	.967
Gore/Dole	1 (59.9%)	1 (59.5%)	.961	.964
Ordered				
Presidential approval	1.266	1.010	1.551	1.294
Age	2.627	2.504	0.990	0.847
Church attendance	1.608	1.322	1.245	1.243
Education	2.204	2.575	0.822	0.626
Income	4.504	5.361	2.185	1.862
Party identification	0.080	0.051	0.901	0.922
Ideology	0.161	0.040	0.911	0.926
Abortion	0.170	0.123	0.490	0.500
Kosovo	–0.082	0.034	1.595	1.537
Newspaper read	1.062	0.804	1.395	1.247
Foreign performance	1.028	0.905	1.480	1.231
Economic performance	1.569	1.314	1.577	1.381

NOTE: Different statistics for central tendency and dispersion were used depending on how the variable was quantified. For categorical variables, we report the mode and the percentage of cases falling in the modal category as measures of central tendency and the index of qualitative variation (Bohrnstedt & Knoke, 1982, p.76) as a measure of dispersion. For ordered variables, we report the mean as a measure of central tendency and the standard deviation as a measure of dispersion.

Internet sample to the telephone sample, we initially conducted a multivariate parameter test (see Table 2, column 2). This test enabled us to determine if the parameter estimates varied as a group by sampling mode. The test of the 14 parameters yielded a chi square of 13.51 ($p >$ .49), indicating that as a whole, the parameters do not significantly vary from Internet users to the general population.

Although the joint test indicated that the parameters did not vary, there could still be a number of individual parameters that differed between the two samples. To test this, we included each interaction term into the model separately (see Table 2, column 3). A significant relationship for the interaction term would indicate that the parameter differed for that particular variable across sampling techniques. After separately entering and testing 14 different interaction terms, only two of these interaction terms passed our threshold for statistical significance—income at the .05 level and marital status at the .10 level.

Vote Choice Models

The results of the vote choice models mirrored closely the results of the approval model, with even less variation occurring across the two samples. The two baseline models, reported in the first columns of Tables 3 and 4, suggested similar explanatory factors.[6] The electoral decision between Bush and Gore was a function of partisanship, presidential approval, ideol-

TABLE 2
Test of the Impact of Sampling Mode on Presidential Approval

	Baseline Model	*Joint Test of Interaction Parameters*	*Individual Test of Interaction Parameters*
Age	–.056	.150	.152
	(.091)	(.188)	(.168)
Gender	–.010	.187	.010
	(.153)	(.320)	(.304)
Married	.012	–.421	–.582*
	(.177)	(.362)	(.315)
Child	–.019	–.260	–.470
	(.164)	(.333)	(.269)
Church attendance	.071	.012	–.016
	(.062)	(.127)	(.122)
Education	–.223**	.070	–.004
	(.104)	(.227)	(.214)
Income	–.045	–.122	–.166**
	(.041)	(.084)	(.076)
Party identification	–.580**	.243	.093
	(.110)	(.235)	(.165)
Ideology	–.037	–.080	–.006
	(.101)	(.212)	(.161)
Abortion	.347*	.166	.300
	(.179)	(.365)	(.293)
Kosovo	–.001	.064	.073
	(.062)	(.129)	(.094)
Newspaper read	–.115**	–.222*	–.184
	(.058)	(.124)	(.117)
Foreign performance	1.047**	.017	–.004
	(.088)	(.168)	(.126)
Economic performance	.794**	–.064	–.060
	(.081)	(.157)	(.116)
Joint χ^2 ($df = 14$)		13.51	
Model χ^2 ($df = 14$)	818.62**		
N	867		

NOTE: Cells in the first column report the ordered logit coefficients for the baseline equation. Cells in the second column report the ordered logit coefficients for the multiplicative terms interacting the baseline variables with the sampling mode from an equation also containing the variables from the baseline model and a dummy variable for the sampling mode. The joint χ^2 reported at the bottom of the second column simultaneously tests the significance of all interaction terms. Cells in the third column report the ordered logit coefficents for each multiplicative term estimated separately with the baseline variables and the mode dummy. If individual χ^2 tests rather than the *z* values of the interaction terms had been used to judge the effect of sampling mode, the reported significance levels would be the same. The coefficients for the additive terms and the dummy variables from the second and third columns were excluded for space considerations but are available upon request. Standard errors for each coefficient reported are in parentheses.
**$p < .05$. *$p < .10$ (two tailed test).

ogy, newspaper readership, education, and abortion attitudes. Alternatively, the choice between Dole and Gore rested on partisanship, presidential approval, ideology, and gender.

Statistical tests demonstrate that the Internet and telephone samples generated similar parameter estimates for both electoral models. The joint chi-square test of the 13 parameters for the Bush/Gore model (see Table 3, column 2) did not achieve statistical significance ($p <$

TABLE 3
Test of the Effect of Sampling Mode on Vote Choice Between G. Bush and A. Gore

	Baseline Model	*Joint Test of Interaction Parameters*	*Individual Test of Interaction Parameters*
Constant	1.385**		
	(.670)		
Age	.174	–.061	–.161
	(.142)	(.319)	(.284)
Gender	.153	.870	.834
	(.251)	(.563)	(.546)
Married	–.286	.149	.008
	(.285)	(.657)	(.533)
Child	.316	.285	.374
	(.268)	(.569)	(.507)
Church attendance	.050	–.312	–.243
	(.105)	(.228)	(.216)
Education	–.292*	.220	.135
	(.165)	(.393)	(.364)
Income	.025	–.060	–.051
	(.066)	(.159)	(.133)
Party identification	1.642**	.204	–.027
	(.157)	(.361)	(.307)
Ideology	.376**	–.351	–.406
	(.147)	(.325)	(.291)
Abortion	.784**	.392	.096
	(.302)	(.645)	(.496)
Kosovo	–.065	.350	.223
	(.101)	(.219)	(.164)
Newspaper read	.265**	.136	.125
	(.010)	(.233)	(.222)
Presidential approval	–.707**	–.071	.009
	(.111)	(.237)	(.214)
Joint χ^2 (df = 13)		9.07	
Model χ^2 (df = 13)	539.37**		
N	763		

NOTE: Cells in the first column report the binary logit coefficients for the baseline equation. Cells in the second column report the binary logit coefficients for the multiplicative terms interacting the baseline variables with the sampling mode from an equation also containing the variables from the baseline model and a dummy variable for the sampling mode. The joint χ^2 reported at the bottom of the second column simultaneously tests the significance of all interaction terms. Cells in the third column report the binary logit coefficients for each multiplicative term estimated separately with the baseline variables and the mode dummy. If individual χ^2 tests rather than the z values of the interaction terms had been used to judge the effect of sampling mode, the reported significance levels would be the same. The coefficients for the additive terms and the dummy variables from the second and third columns were excluded for space considerations but are available upon request. Standard errors for each coefficient reported are in parentheses.
**$p < .05$. *$p < .10$ (two-tailed test).

.76), supporting the notion of the overall equivalence of parameters between the samples. Moreover, when the interaction terms were individually entered into the model (see Table 3, column 3), none of the interaction parameters achieved statistical significance, meaning that the decision-making process that produced electoral support for Bush or Gore was the same for Internet users and the general population.

TABLE 4
Test of the Effect of Sampling Mode on Vote Choice Between E. Dole and A. Gore

	Baseline Model	*Joint Test of Interaction Parameters*	*Individual Test of Interaction Parameters*
Constant	.250		
	(.599)		
Age	.096	.289	.303
	(.129)	(.287)	(.248)
Gender	.486**	–.308	–.285
	(.221)	(.488)	(.440)
Married	–.346	.360	.227
	(.249)	(.547)	(.448)
Child	.164	–.452	–.469
	(.235)	(.493)	(.432)
Church attendance	.056	–.074	–.106
	(.089)	(.190)	(.177)
Education	.119	–.152	.118
	(.143)	(.344)	(.305)
Income	.065	.046	.044
	(.056)	(.127)	(.109)
Party identification	1.213**	–.087	–.283
	(.139)	(.321)	(.257)
Ideology	.551**	–.785**	–.753**
	(.125)	(.275)	(.251)
Abortion	.055	–.460	–.078
	(.259)	(.547)	(.422)
Kosovo	.047	.234	.157
	(.087)	(.192)	(.140)
Newspaper read	.037	–.215	–.242
	(.086)	(.208)	(.188)
Presidential approval	–.546 **	–.235	–.063
	(.095)	(.208)	(.176)
Joint χ^2 (df = 13)		16.05	
Model χ^2 (df = 13)	398.28**		
N	745		

NOTE: Cells in the first column report the binary logit coefficients for the baseline equation. Cells in the second column report the binary logit coefficients for the multiplicative terms interacting the baseline variables with the sampling mode from an equation also containing the variables from the baseline model and a dummy variable for the sampling mode. The joint χ^2 reported at the bottom of the second column simultaneously tests the significance of all interaction terms. Cells in the third column report the binary logit coefficents for each multiplicative term estimated separately with the baseline variables and the mode dummy. If individual χ^2 tests rather than the *z* values of the interaction terms had been used to judge the effect of sampling mode, the reported significance levels would be the same. The coefficients for the additive terms and the dummy variables from the second and third columns were excluded for space considerations but are available upon request. Standard errors for each coefficient reported are in parentheses.
**$p < .05$. *$p < .10$ (two-tailed test).

The model of the Dole/Gore decision yielded similar results. The joint chi-square test (Table 4, column 2) of all interaction parameters failed to meet standard thresholds of significance ($p > .24$), suggesting that the parameters derived from each sampling mode do not vary. Moreover, when the interaction terms enter the model individually (see Table 4, column 3), only the effect of ideology differed across the samples ($p < .05$).[7]

CONCLUSIONS

Overall, our analysis suggests that Internet samples may be useful for investigating how individuals generate certain types of attitudes. Internet users and nonusers seem to use similar psychological mechanisms to arrive at common political decisions such as vote choice and candidate performance evaluation. In other words, we would have reached the same conclusions about the determinants of particular political attitudes by relying on a diverse convenience sample of Internet users as we would have by using a more expensive, time-consuming, probabilistic telephone sample.

Nevertheless, before Internet samples are used more pervasively to analyze mass behavior, several important concerns remain to be addressed. First, researchers should consider how widespread the similarities are between the decision-making processes of Internet users and nonusers. Although we found evidence of parallel psychological mechanisms for political candidate evaluations, there are no assurances that decision-making processes underlying other judgments, political or nonpolitical, will be similar. For instance, it is unlikely that Internet users and nonusers draw on similar factors to generate preferences toward technology policy, feelings of social connectedness, or beliefs about the trustworthiness of various news sources. Any explanatory factor related to the Internet experience becomes particularly vulnerable to producing different causal mechanisms for Internet users and nonusers. Scholars should be wary of relying exclusively on Internet samples until the effects of Internet use are more fully understood.

Second, researchers should investigate the means by which diverse samples are drawn. Because we currently do not have the ability to draw representative samples of Internet users, even if the decision-making processes for Internet users and nonusers were the same, we must devise a sampling technique that consistently produces diverse samples. Although we generated sufficient sample diversity in this analysis by drawing e-mail addresses from public service directories, this may have been a fortunate consequence of the timing of our survey and the hypotheses we tested. Research should examine how choices about sampling units, sampling selection, and sample design influence the dispersion of common measures before scholars begin haphazardly generating samples of Internet users.

Finally, researchers should examine how Internet accessibility influences decision making. Regardless of the sampling approach employed, numerous Internet users will be excluded from the sample because they are inaccessible by the technique. This raises several important concerns. Is accessibility a function of particular factors or is it merely a coincidence? Does accessibility mediate or moderate how hypothesized models operate? If so, when? Without understanding these questions, we cannot rely on diverse samples, even in those instances where we are confident that the mechanisms underlying Internet user are the same as nonusers.

The Internet is an exciting new resource that offers unique research opportunities. However, until every individual in the population has access and is accessible by it, our ability to infer the nature of the population from analyses of Internet users hinges on a series of assumptions. Although this article demonstrates limited support for some of these assumptions, considerable research remains to be done before researchers can rely on Internet samples without hesitation.

NOTES

1. Even among accessible Internet users, systematic differences may emerge as a result of the lack of synchrony between interviewer and interviewee. Some users take time to contemplate their responses whereas others respond immediately. Some users respond spontaneously whereas others deliberate their responses.

2. Such a technique eliminates roughly half of the nonresidential, nonworking numbers, saving considerable time and resources.

3. Interviewing times were slightly different on the weekend. Telephone calls were conducted on Saturdays from 10 a.m. to 6 p.m. and on Sundays from 12 noon to 8 p.m.

4. The following question wordings were used to assess the respondents' demographics. Age: "In what year were you born?" Gender: "What is your gender, male or female?" Church attendance: "How often do you attend religious services apart from occasional weddings, baptisms, or funerals: more than once a week, once a week, once or twice a month, a few times a year, or never?" Education: "What is the highest grade in school or level of education that you've completed and got credit for?" Marital status: "Are you currently married, widowed, divorced, separated, or have you never been married?" Children: "How many of the persons who currently live in your household are under 18 years of age, including babies and small children?" Income: "How much income did you receive in 1998, not just from wages or salaries but from all sources—that is, before taxes and other deductions were made? Was it less than $20,000, $20,000-$39,999, $40,000-$59,999, $60,000-$79,999, $80,000-$99,999, $100,000 and over?" Newspaper readership: "How often do you read the daily or Sunday edition of the *Manchester Union Leader* newspaper: every day, almost every day, several times a week, occasionally, or never?" Race was omitted from the survey because non-Whites make up less than 2% of New Hampshire residents.

5. To estimate the models, we used Intercooled Stata 6, Stata Corporation, 702 University Drive East, College Station, TX 77840; e-mail: `stata@stata.com`

6. We found no statistical difference in the distribution of demographics for the different presidential pairing of candidates within both the e-mail and telephone samples.

7. The results of the individual parameter tests are even more convincing when considered across the three models. Given a $p < .10$ significance threshold and 40 individual parameter tests, we would expect up to four deviations based on chance alone.

REFERENCES

Abramowitz, A. I. (1995). It's abortion, stupid: Policy voting in the 1992 presidential election. *The Journal of Politics*, *57*, 176-186.

Anderson, C. A., Lindsay, J. J., & Bushman, B. J. (1999). Research in the psychological laboratory: Truth or triviality? *Current Directions in Psychological Science*, *8*, 3-9.

Aquilino, W. S. (1994). Interview mode effects in surveys of drug and alcohol use. *Public Opinion Quarterly*, *58*, 210-240.

Babbie, E. R. (1990). *Survey research methods* (2nd ed.). New York: Wadsworth.

Bimber, B. (1998). The Internet and political mobilization: Research note on the 1996 election season. *Social Science Computer Review*, *16*, 391-401.

Binik, Y. M., Mah, K., & Kiesler, S. (1999). Ethical issues in conducting sex research on the Internet. *The Journal of Sex Research*, *36*, 82-90.

Bohrnstedt, G. W., & Knoke, D. (1982). *Statistics for Social Data Analysis*. Itasca, IL: F.E. Peacock Publishers.

Bonchek, M. S., Hurwitz, R., & Mallery, J. (1996). Will the Web democratize or polarize the political process? *World Wide Web Journal*, *3*, 1-6.

Buchanan, T., & Smith, J. L. (1999). Using the Internet for psychological research: Personality testing on the World Wide Web. *British Journal of Psychology*, *90*, 125-144

Davis, R. N. (1999). Web-based administration of a personality questionnaire: Comparison with traditional methods. *Behavior Research Methods, Instruments, and Computers 31*, 572-577.

De Leeuw, E. D., Mellenbergh, G. J., & Hox, J. J. (1996). The influence of data collection method on structural models. *Sociological Methods & Research*, *24*, 443-472.

Fisher, B., Margolis, M., & Resnick, D. (1996). Surveying the Internet: Democratic theory and civic life in cyberspace. *Southeastern Political Review*, *24*, 399-429.

Georgia Institute of Technology, Graphic, Visualization, & Usability Center. (1998). *GVU's Tenth WWW User Survey Report* [Online]. Available: `http://www.cc.gatech.edu/gvu/user_surveys/`

Gilens, M. (1988). Gender and support for Reagan: A comprehensive model of presidential approval. *American Journal of Political Science*, *32*, 19-49.

Gotcher, J. M., & Kanervo, E. W. (1997). Perceptions and uses of electronic mail. *Social Science Computer Review*, *15*, 145-158.

Hewson, C. M., Laurent, D., & Vogel, C. M. (1996). Proper methodologies for psychological and sociological studies conducted via the Internet. *Behavior Research Methods, Instruments, and Computers*, *28*, 186-191.

Hurwitz, J., & Peffley, M. (1987). The means and ends of foreign policy as determinants of presidential support. *American Journal of Political Science*, *31*, 236-258.

Johnson, T. J., & Kaye, B. K. (1998). A vehicle for engagement or a haven for the disaffected? Internet use, political alienation, and voter participation. In T. J. Johnson, C. E. Hays, & S. P. Hays (Eds.), *Engaging the public: How government and the media can reinvigorate American democracy* (pp. 123-135). New York: Roman & Littlefield.

Judge, G. G., Hill, R. C., Griffiths, W. E., Lütkepohl, H., & Lee, T. C. (1988). *Introduction to the theory and practice of econometrics* (2nd ed.). New York: Wiley.

Krantz, J. H., Ballard, J., & Scher, J. (1997). Comparing the results of laboratory and World-Wide Web samples on the determinants of female attractiveness. *Behavior Research Methods, Instruments, and Computers*, *29*, 264-269.

Krosnick, J. A., & Alwin, D. F. (1988). A test of the form-resistant correlation hypothesis: Ratings, rankings, and the measurement of values. *Public Opinion Quarterly*, *52*, 526-538.

Lang, A. (1996). The logic of using inferential statistics with experimental data from nonprobability samples: Inspired by Cooper, Dupagne, Potter, and Sparks. *Journal of Broadcasting & Electronic Media*, *40*, 422-430.

Maddala, G. S. (1992). *Introduction to econometrics* (2nd ed.). New York: Macmillan.

Mavis, B. E., & Brocato, J. J. (1998). Postal surveys versus electronic mail surveys: The tortoise and the hare revisited. *Evaluation & The Health Professions*, *21*, 395-408.

National Telecommunications & Information Administration. (2000). *Falling through the net: Toward digital inclusion* [Online]. Available: `http://www.ntia.doc.gov/ntiahome/fttn00/contents00.html`

Nielsen//NetRatings. (2000). *Internet usage statistics for the week of December 10, 2000 [Online]. Available: `http://209.249.142.27/nnpm/owa/nrpublicreports.usageweekly`*

Schaefer, D. R., & Dillman, D. A. (1998). Development of a standard e-mail methodology: Results of an experiment. *Public Opinion Quarterly*, *62*, 378-397.

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, *51*, 515-530.

Shanks, J. M., and Miller, W. E. (1990). Partisanship, policy and performance: The Reagan legacy in the 1988 election. *British Journal of Political Science*, *21*, 129-197.

Schmidt, W. C. (1997). World-Wide Web survey research: Benefits, potential problems, and solutions. *Behavior Research Methods, Instruments, & Computers*, *29*, 274-279.

Smith, M. A., & Leigh, B. (1997). Virtual subjects: Using the Internet as an alternative source of subjects and research environment. *Behavior Research Methods, Instruments, and Computers*, *29*, 496-505.

Sproull, L. S. (1986). Using electronic mail for data collection in organizational research. *Academy of Management Journal*, *29*, 159-169.

Stern, S. E., & Faber, J. E. (1997). The lost e-mail method: Milgram's lost-letter technique in the age of the Internet. *Behavior Research Methods, Instruments, and Computers*, *29*, 260-263.

Swoboda, W. J., Mühlberger, N., Weitkunat, R., & Schneeweib, S. (1997). Internet surveys by direct mailing: An innovative way of collecting data. *Social Science Computer Review*, *15*, 242-255.

Wherrett, J. R. (1999). Issues in using the Internet as a medium for landscape preference research. *Landscape and Urban Planning*, *45*, 209-217.

Witte, J. C., Amoroso, L. M., & Howard, P. E. (2000). Research methodology: Method and representation in Internet-based survey tools—mobility, community, and cultural identity in Survey2000. *Social Science Computer Review*, *18*, 179-195.

Samuel J. Best is an assistant professor of political science at the University of Connecticut, 341 Mansfield Road, Storrs, CT 06269-1024.

Brian Krueger is a graduate student in the Department of Government at the University of Notre Dame, Notre Dame, IN 46556-0368.

Clark Hubbard is an assistant professor of political science at the University of New Hampshire, Durham, NH 03824-3586.

Andrew Smith is the director of the University of New Hampshire Survey Center, Thompson Hall Room G09, 105 Main Street, Durham, NH 03824.