National Surveys Via RDD Telephone Interviewing vs. the Internet:

Comparing Sample Representativeness and Response Quality

LinChiat Chang   and   Jon A. Krosnick
TNS Global   Stanford University

(corresponding author)

Contact info: Contact info:

3118 18th Street   434 McClatchy Hall
San Francisco, CA 94110   Stanford, CA 94305

Phone. (415) 863-3717   Phone. (650) 725-3031
Fax. n/a   Fax. (650) 725-2472
Email. linchiat.chang@tns-global.com   Email. krosnick@stanford.edu

December, 2008

Acknowledgements & Disclosures

RDD Telephone vs. Internet Survey Methodology:

Comparing Sample Representativeness and Response Quality

Abstract

In a national field experiment, the same questionnaires were administered simultaneously by RDD telephone interviewing, by the Internet with a probability sample, and by the Internet with a non-probability sample of people who volunteered to do surveys for money. The probability samples were more representative of the nation than the non-probability sample in terms of demographics and electoral participation, even after weighting. The non-probability sample was biased toward being highly engaged in and knowledgeable about the survey's topic (politics). The telephone data manifested more random measurement error, more survey satisficing, and more social desirability response bias than did the Internet data, and the probability Internet sample manifested more random error and satisficing than did the volunteer Internet sample. Practice at completing surveys increased reporting accuracy among the probability Internet sample, and deciding only to do surveys of on topics of personal interest enhanced reporting accuracy in the non-probability Internet sample. Thus, the non-probability Internet method yielded the most accurate self-reports from the most biased sample, while the probability Internet sample manifested the optimal combination of sample composition accuracy and self-report accuracy. In a laboratory experiment, respondents were randomly assigned to answer questions either on a computer or over an intercom system with an interviewer; the former respondents manifested higher concurrent validity, less survey satisficing, and less social desirability response bias. These results suggest that Internet data collection from a probability sample yields more accurate results than do telephone interviewing and Internet data collection from non-probability samples.

RDD Telephone vs. Internet Survey Methodology:

Comparing Sample Representativeness and Response Quality

During the history of survey research, the field has witnessed many transitions in the uses of various modes of data collection for interviewing nationally representative samples of adults. Initially, face-to-face interviewing was the predominant method of data collection, yielding high response rates, permitting the development of good rapport between interviewers and respondents, and allowing the use of visual aids that facilitated the measurement process. But the cost of face-to-face interviewing increased dramatically since the 1970s (De Leeuw and Collins 1997; Rossi, Wright, and Anderson 1983), prompting researchers to explore alternative methods, such as telephone interviewing, self-administered paper-and-pencil mail questionnaires (Dillman 1978), audio computer-assisted self-interviewing (ACASI), telephone audio computer-assisted self-interviewing (T-ACASI), Interactive Voice Response (IVR) surveys (Dillman 2000), and more. Among these alternatives, telephone interviewing of samples generated by random digit dialing became a very popular method during the last 30 years, an approach encouraged by studies done in the 1970s suggesting that telephone data quality was comparable to that obtained from face-to-face interviews (e.g., Groves and Kahn 1979).

Recent years have seen a surge in the challenges posed by telephone interviewing. It has become increasingly difficult to maintain response rates, causing the costs of data collection to rise considerably. It is possible to achieve response rates nearly as high as those observed 20 years ago, but doing so costs a great deal more. So holding budget constant over time, the response rate that can be obtained today is considerably lower than that which was obtainable in 1980 (Holbrook, Krosnick, and Pfent 2007; Lavrakas 1997).

Against this backdrop, Internet surveys appeared as a promising alternative about ten years ago. Some survey firms that had concentrated their efforts on telephone interviewing shifted to collecting a great deal of data over the Internet, including the Gallup Organization and Harris Interactive. And

other firms were newly created to take advantage of the Internet as a data collection medium, including Knowledge Networks and Greenfield Online.

Resistance to new modes of data collection is nothing new in the history of survey research, and it is as apparent today as it has been in the past. Just as the survey industry was reluctant to embrace the telephone when it emerged decades ago as an alternative to face-to-face interviewing, some researchers today are hesitant about a shift to internet-based data collection when the goal is to yield representative national samples. This skepticism has some basis in reality: there are notable differences between face-to-face, telephone, and mail surveys on the one hand and Internet surveys on the other in terms of sampling and recruitment methods, most of which justify uncertainty about the quality of data obtained via Internet surveys (e.g., Couper 2000).

Nonetheless, practical advantages of Internet surveys are obvious. Computerized questionnaires can be distributed easily and quickly via web sites postings or hyperlinks or attachments to emails. No travel costs, postage or telephone charges, or interviewer costs are incurred. Respondents can complete questionnaires on their own whenever it is convenient for them. Turn-around time can be kept short, and the medium allows easy presentation of complex visual and audio materials to respondents, implementation of complex skip patterns, and consistent delivery of questions and collection of responses from respondents. Therefore, it is easy to understand why many survey practitioners today find the Internet approach potentially appealing in principle. But for the serious research community, practical conveniences are of limited value if a new methodology brings with it declines in data quality. Therefore, to help the field come to understand the costs and benefits of Internet data collection, we initiated a project to compare this method to one of its main competitors: telephone surveying.

We begin below by outlining past mode comparison studies and compare Internet and telephone survey methodologies in terms of potential advantages and disadvantages. Next, we describe the design of a national field experiment comparing data collected by Harris Interactive (HI), Knowledge Networks (KN), and the Ohio State University Center for Survey Research (CSR) and

detail the findings of analyses comparing sample representativeness and response quality. In Appendix 3, we report a supplementary laboratory experiment comparing the modes as well.

Past Research Involving Mode Comparisons

There is a long tradition of mode comparison studies in survey research. Many studies have compared face-to-face surveys with telephone surveys (e.g., de Leeuw and Van der Zouwen 1988; Holbrook, Green, and Krosnick 2003) and paper and pencil surveys (e.g., de Leeuw 1992; Hox and de Leeuw 1994), and others have compared telephone surveys with paper and pencil surveys (e.g., de Leeuw 1992; Gano-Phillips and Fincham 1992). Comparing telephone surveys with Internet surveys is a much newer activity, generating a growing number of studies (e.g., Berrens, Bohara, Jenkins-Smith, Silva, and Weimer 2003; Fricker, Galesic, Tourangeau, and Yan 2005).

Past mode comparisons have examined a variety of dependent measures, including response rates and cost (e.g., Fournier and Kovess 1993), demographic distributions (e.g., Aneshensel, Frerichs, Clark, and Yokopenic 1982), distributions of key response variables (e.g., Rockwood, Sangster, and Dillman 1997), item nonresponse (e.g., Herzog and Rodgers 1988), and self-reports on sensitive topics (e.g., Sudman, Greeley, and Pinto 1965). Other endeavors have compared response biases (e.g., Tarnai and Dillman 1992), response stability across time (e.g., Martin, O'Muircheartagh, and Curtis 1993), pace of interviewing (e.g., Groves 1978), and structural equation model parameter estimates (e.g., de Leeuw, Mellenbergh, and Hox 1996) between modes.

Sometimes, data collected via different modes have been compared against independent, objective criteria (e.g., Biemer 2001; Siemiatycki 1979). For example, demographic distributions of samples from telephone vs. face-to-face surveys have been evaluated by how well they matched the estimates from the Current Population Survey (the CPS; Holbrook, Green, and Krosnick 2003). Following in this tradition, the present investigation assessed sample representativeness by comparing demographic distributions to benchmarks obtained from the CPS. We also studied mode differences in

distributions of key response variables and assessed data quality using regression coefficients and structural equation model parameter estimates.

Internet and Telephone Survey Methodologies

Two primary methodologies have been employed by commercial firms conducting surveys via the Internet. One method, employed by companies such as Harris Interactive (HI), begins by recruiting potential respondents through invitations that are widely distributed in ways designed to yield responses from heterogeneous population subgroups with Internet access. Another approach entails probability samples reached via Random Digit Dialing who are invited to join an Internet survey panel; people without computer equipment or Internet access are given it at no cost. Two firms taking this approach in the U.S. (Knowledge Networks (KN) and the RAND Corporation) and have given WebTV or MSNTV equipment and service to respondents who needed them. These two approaches to recruit panel members have been outlined in detail elsewhere (Berrens et al. 2003; Best et al. 2001; Chang 2001; Couper 2000), so we describe briefly the methods used at the time when our data collection occurred.

Harris Interactive Methodology. Harris Interactive recruited more than three-quarters of their panel members from one of the most popular Internet search engines: www.excite.com. On the main page of excite.com, a link appeared inviting visitors to participate in the poll of the day. Respondents who voted on the day's issue then saw a link inviting them to become panel members for the Harris Poll Online (HPOL). The second source of panel members was the website of Matchlogic, Inc., an online marketing company and a subsidiary of Excite@Home. Matchlogic posted banner advertisements on the Internet to attract consumers with promises of incentives such as free merchandise and sweepstakes. When a person registered for those incentives, he or she was invited to become a panel member for HPOL. At the time when our data collection occurred, Excite and Matchlogic accounted for about 90% of all panel members; the others were recruited by invitations on other websites.

People visiting the Harris Poll Online (HPOL) registration site were asked for their email addresses and demographic information and were told that HPOL membership would allow them to influence important decision-makers in government, non-profit organizations, and corporations, could help to shape policies, products, and services, would have access to some survey results prior to publication in the media, and might win cash, free consumer products, or discount coupons, or receive other tangible incentives.

Harris Interactive's database has contained more than 7 million panel members, and subsets of these individuals were invited to participate in particular surveys. A panel member who was invited to do a survey could not be invited to do another survey for at least 10 days, and each panel member usually received an invitation at least once every few months. Survey completion rates ranged from a low of 5% to a high of 70%, with an average of 15-20% (R. Thomas, personal communication).

To generate a national sample, panel members were selected based on demographic attributes (e.g., age, gender, region of residence) so that the distributions of these attributes in the final sample matched those in the general population. Each selected panel member was sent an email invitation that described the content of the survey and provided a hyperlink to the website where the survey was posted and a unique password allowing access to the survey. Respondents who did not respond to an email invitation or did not finish an incomplete questionnaire were sent reminder emails.

Harris Interactive weighted each sample using demographic data from the Current Population Survey (CPS) and answers to questions administered in Harris Poll monthly telephone surveys of national cross-sectional samples of 1,000 American adults, aged 18 and older. The goal of their weighting procedure was to adjust for variable propensity of individuals to have regular access to email and the Internet to yield results that can be generalized to the general population.

Knowledge Networks Methodology. Beginning in 1998, Knowledge Networks recruited panel members through RDD telephone surveys and provided people with WebTV equipment and Internet access in exchange for their participation in surveys. Knowledge Networks excluded telephone

numbers from their RDD samples that were not in the service area of a WebTV Internet service provider, leading to exclusion of about 6-7% of the general population. Knowledge Networks attempted to obtain mailing addresses for all sampled telephone numbers and succeeded in doing so for about 60% of them. These households were then sent advance letters stating that they had been selected to participate in a survey panel, that they would not pay any cost, that confidentiality was assured, and that a Knowledge Networks staff member would call them within a week. A $5 or $10 bill was included with the letter to encourage cooperation.

Telephone interviews were attempted with all households that received an advance letter. Telephone interviews were also attempted with one-third (selected randomly) of the telephone numbers for which an address could not be obtained. During the telephone interviews, respondents were told they had been selected to participate in an important national study using a new technology and that they would be given a WebTV receiver that would allow them free access to the Internet and opportunities to answer brief surveys on their television. Respondents were told that their participation was important and were asked about the extent to which household members were experienced with the Internet and proficient with computers and about some demographics of household members. Arrangements were then made to mail the WebTV equipment to the respondent.

After households received the WebTV equipment and installed it (with assistance from a technical support office via telephone if necessary), respondents completed "profile" surveys that measured many attributes of each adult household member. Each adult was given a free email account and was asked to complete surveys via WebTV. Whenever any household member had a new email message waiting to be read, a notification light flashed on the WebTV receiver (a box located near the television set). Panel members could then log into their WebTV accounts and read the email invitation for the survey, which contained a hyperlink to the questionnaire. Panel members were usually sent one short survey per week, typically not exceeding 15 minutes. When a panel member was asked to

complete to a longer questionnaire, he or she was then given a week off or offered some other form of incentive or compensation.

Typically, about 85% of respondents who were asked to complete a KN survey did so within 2 weeks of the invitation, and few responses were received after that. Respondents who failed to respond to eight consecutive surveys were dropped from the panel, and the WebTV receiver was removed from their homes.

Thus, households intended to provide data for any given survey could fail to do so because of dropout at several stages throughout the recruitment process. Some households were excluded because they did not live in an area covered by a WebTV provider. Some households were not contacted to complete the initial RDD telephone interview. Other households were contacted but refused to join the panel. Of the households that signed up initially, some failed to install the WebTV equipment in their homes. And some people who had the equipment installed either failed to complete a questionnaire for a particular survey or dropped out of the panel altogether after joining it.

COVERAGE AND NON-RESPONSE ERROR

HI samples entailed coverage error, because they include only people who had pre-existing access to computers and the Internet, thus probably over-representing urban residents, men, wealthier, more educated, younger, and White people (Flemming and Sonner 1999; Rohde and Shapiro 2000). Although KN provided Internet access to its panel members, its sampling technique brought with it the same coverage error inherent in all RDD telephone surveys, excluding about 5% of the country's population because their households were without working telephones. If respondents who already had Internet access in their homes were more likely to reject the offer of free Internet access via WebTV, then the KN samples would under-represent regular Internet users.

TOPIC INTEREST

The method used by most non-probability Internet firms to invite respondents to complete a questionnaire may create sample composition bias driven by interest in the topic. In the email

invitations sent to selected HI respondents, a one-sentence description informed people about the content of the survey (e.g., telecommunications, entertainment, or politics). People interested in the topic may have been more likely to participate than people not so interested. Although the HI weighting procedure adjusts for demographic attributes and other variables, the adjustment procedure have not usually corrected for interest in a particular topic.

ADVANTAGES OF THE TELEPHONE OVER THE INTERNET

There are many reasons why the quality of survey responses collected via the Internet might differ from those collected by telephone. One potential strength of telephone surveying is the presence of interviewers, who can provide positive feedback to respondents in order to encourage effortful engagement in the response process (Cannell, Miller, and Oksenberg 1981). Likewise, interviewers can project interest and enthusiasm, which may be unconsciously contagious (Chartrand and Bargh 1999), and respondents' moods can be unconsciously enhanced by the emotions in the interviewer's voice (Neumann and Strack 2000). Thus, if interviewers' voices transmit interest and enthusiasm about a survey, they may inspire increased respondent engagement. Such processes cannot occur when respondents complete self-administered questionnaires.

Interviewers can also create a sense of accountability among respondents due to "the implicit or explicit expectation that one may be called on to justify one's beliefs, feelings, and actions to others" (Lerner and Tetlock 1999). In past research, participants who reported their judgments aloud to another person recognized that their judgments were linked directly to them in the eyes of the individual with whom they were interacting, resulting in high accountability (e.g., Lerner, Goldberg, and Tetlock 1998; Price 1987). When the audience's views on the issues in question are not known, accountability generally leads people to devote more careful and unbiased effort to making judgments (for a review, see Lerner and Tetlock 1999). Although survey responding via the Internet to HI and KN is not anonymous, the palpable phenomenology of accountability under those circumstances may be considerably less than when a respondent is conversing with an interviewer. Therefore, this may

increase the precision of survey responses provided over the telephone.

Telephone surveys have another potential advantage over Internet surveys: respondents do not need to be literate or be able to see clearly enough to read words, because all questions and answer choices are read aloud to them. Telephone respondents also do not need to be proficient at using a computer or to be knowledgeable about how to navigate the Internet. Thus, telephone surveys may be more manageable than Internet surveys.

ADVANTAGES OF THE INTERNET OVER THE TELEPHONE

Just as interviewers may be advantageous, they may also entail drawbacks, so the absence of interviewers might be a strength of Internet surveys. Interviewers are known to create errors and biases when collecting data (Kiecker and Nelson 1996). Due to misunderstandings, bad habits, or biased expectations, some interviewers occasionally provide inappropriate cues (van der Zouwen, Dijkstra, and Smit 1991) or change the wordings of questions (Lyberg and Kasprzyk 1991). None of this can occur in an Internet survey.

Some studies suggest that people are more concerned about presenting a favorable self-image during oral interviews than when completing self-administered questionnaires (Acree, Ekstrand, Coates, and Stall 1999; Fowler, Roman, and Di 1998). If self-administered questionnaires do indeed decrease concern about impression management, people may be less likely to conform to socially desirable standards and more likely to provide honest answers to questions on threatening, anxiety-arousing, or otherwise sensitive questions (e.g., Tourangeau and Smith 1996; Wright, Aquilino, and Supple 1998).

Pauses can feel awkward during telephone conversations, which may induce interviewers and respondents alike to rush the speed of their speech, making it difficult for respondents to understand questions and to calmly reflect on the meaning of a question or think carefully to generate an accurate answer. In contrast, Internet respondents can set their own pace when completing a survey, pausing to deliberate about complex questions and moving quickly through questions that are easy to interpret and answer. In addition, Internet respondents can take breaks when they are fatigued and return refreshed.

These factors may facilitate better efficiency and precision in answering by Internet respondents.

Internet respondents have the flexibility to complete a questionnaire at any time of day or night that is convenient for them. Telephone interviewing organizations allow for call scheduling at times that are convenient for respondents, but their flexibility in call scheduling falls short of the 24-hour accessibility of Internet surveys. Thus, Internet respondents can choose to complete a survey when they are most motivated and able to do so and when distractions are minimized, perhaps causing improved response quality.

Telephone respondents need to hold a question and its response options in working memory in order to answer accurately. Because Internet respondents can see questions and response categories, they need not commit them to memory before generating answers. If respondents fail to remember the details of a question after reading it once, they can read the question again. And when long checklists or complex response scales are used, Internet respondents are not especially challenged, because the response options are fully displayed. This may reduce the cognitive burden of the response process and may thereby improve reporting accuracy.

PRACTICE EFFECTS

RDD telephone surveys typically involve respondents who have some experience responding to questionnaires[1], but KN and HI respondents were members of long-term panels and therefore had regular practice at survey responding. A great deal of psychological research shows that practice at cognitive tasks improves performance on them, so regular experience answering survey questions may enhance the accuracy of Internet panel members' responses (Donovan and Radosevich 1999; Smith, Branscombe, and Bormann 1988). Also, panel members may become especially self-aware and introspective about their thoughts, attitudes, emotions, and behaviors, further improving their ability to later report on those phenomena accurately (Menard 1991). Consistent with this reasoning, research on panel surveys has shown that people's answers to attitude questions become increasingly reliable as they gain more experience responding to them (Jagodzinski, Kuhnel, and Schmidt 1987).

POTENTIAL DRAWBACKS OF INTERNET PANELS

A potential drawback of repeated interviewing is "panel conditioning," whereby accumulating experience at doing surveys makes panel members less and less like the general public they are intended to represent. A number of studies exploring this possibility have found either no evidence of panel conditioning effects or very small effects. For example, Cordell and Rahmel (1962) found that participating in Nielsen surveys on media use did not alter later reports of media use. Likewise, Himmelfarb and Norris (1987) found that being interviewed on a wide range of topics did not alter people's subsequent reports of mental health, physical health, self-esteem, social support, or life events experienced (see also Sobol 1959; Clinton 2001). Willson and Putnam (1982) found in a meta-analysis that answering questions caused attitudes toward objects to become slightly more positive, but these effects were quite small and inconsistent across studies.

Some studies that documented conditioning effects tested the "stimulus hypothesis" (Clausen 1968): the notion that interviewing people on a particular topic may induce them to become more cognitively engaged in that topic subsequently. Some studies found support for this notion (e.g., Bridge, Reeder, Kanouse, Kinder, Nagy, and Judd 1977; Granberg and Holmberg 1991), though others did not (e.g., Mann 2005). Other studies have documented how asking people just one question about their behavioral intentions could impact on subsequent behavior (e.g., Fitzsimons and Morwitz 1996; Sherman 1980; Greenwald, Carnot, Beach, and Young 1987). Thus, this literature clearly suggests that panel conditioning effects can occur (see also the literature on pretest sensitization; e.g., Bracht and Glass 1968).

Another potential drawback of panel studies involve respondent attrition: Some of the people who provide data during the first wave of interviewing do not participate in subsequent waves. If a non-random subset of respondents drop out, then this would compromise sample representativeness. However, the literature on panel attrition is actually quite reassuring on this point. Although a few past studies have documented instances in which people who were and were not reinterviewed differed

14

from one another in some regard (e.g., Groves, Singer, and Corning 2000; Lubin, Levitt, and Zuckerman 1962), most studies found little or no sample composition changes as individual panel members dropped out of the active samples (e.g., Zagorsky and Rhoton 1999; Fitzgerald, Gottschalk, and Moffitt 1998a, 1998b; Falaris and Peters 1998; Clinton 2001) Hence, panel attrition may not be a serious drawback of internet panels, especially when the topics of the survey vary from week to week.

SUMMARY

In sum, if interviewers bring positive reinforcement, enthusiasm, and accountability into the survey process and literacy is a significant problem, then response quality may be advantaged by the presence of interviewers and oral presentation in telephone surveys. But if a greater sense of privacy, self-pacing, flexibility to complete surveys at any time of day or night, an ability to see questions and response options, and practice effects advantage the Internet mode, then response quality may be better in such surveys than in telephone surveys. In the present investigation, we did not set out to explicitly gauge the impact of each of the factors outlined above. Rather, we set out to ascertain whether mode differences existed in sample representativeness and response quality, and the extent and direction of such differences if they existed.

The National Field Experiment

For our study, HI, KN, and the OSU CSR collected data in two waves, once before the 2000 U.S. Presidential election campaign began, and then again after election day. During the pre-election survey, respondents predicted their presidential vote choice in the elections, evaluated the leading presidential candidates, and reported a wide range of attitudes and beliefs that are thought to drive vote choices. During the post-election survey, respondents reported whether they had voted, for whom they had voted, and again evaluated the leading presidential candidates. These questions are documented in Appendix 1.

Our comparisons focused on the demographic composition of the samples in terms of

education, age, race, gender, and income; the samples' interest in the survey topic; concurrent validity of the measures (i.e., their ability to distinguish between people on a criterion measured at the same point in time; e.g., Leary 1995); predictive validity of the measures (i.e., their ability to predict a criterion measured some time in the future; e.g., Aronson, Ellsworth, Carlsmith, and Gonzales 1990; Leary 1995); the extent of survey satisficing (Krosnick 1991, 1999), reliability, and social desirability response bias.

To assess concurrent and predictive validity, we conducted regressions predicting people's pre-election predictions of their vote choice and their post-election reports of their actual vote choices using a plethora of predictors of presidential candidate preferences. The logic underlying these criterion validity analyses is displayed in Figure 1. Each determinant of vote choice (shown in the lower left corner of Figure 1) is expected to be associated with true vote choice (shown in the lower right corner of Figure 1), and the true magnitude of this association is ř. Self-reports of these two constructs appear on the top of Figure 1. By correlating self-reports measured pre-election (shown in the upper left of Figure 1) with post-election reports of vote choice (shown in the upper right of Figure 1), we obtain r_{12}. The lower the validity of the items (represented by λ_1 and λ_2) and the more measurement error in reports (represented by ε_1 and ε_2), the more r_{12} will be weakened in comparison to ř. Thus, the observed strength of the relation between a measure of a vote choice determinant and vote choice is an indicator of response quality. The more respondents are willing and able to precisely report vote choice and its determinants, the stronger the relation between these two manifest variables will presumably be.

To assess survey satisficing, which occurs when respondents do not engage in careful and thorough thinking to generate accurate answers to questions (Krosnick 1991, 1999), we looked at non-differentiation in answering batteries of questions using identical rating scales. Non-differentiation occurs when respondents rate several target persons or issues or objects nearly identically on a single dimension because they do not devote effort to the reporting process. Although a set of identical ratings across objects may be the result of genuinely similar attitudes, non-differentiation tends to occur

under conditions that foster satisficing.

To test whether the data collection methods differed in terms of the amount of random measurement error in assessments, we made use of multiple measures of candidate preferences administered both pre-election and post-election to estimate the parameters of a structural equation model (see, e.g., Kenny 1979). This model posited that the multiple measures were each imperfect indicators of latent candidate preferences at the two time points and permitted those preferences to change between the two interviews.

Finally, we examined whether the two modes differed in terms of social desirability response bias. The survey questionnaire contained a question about whether the federal government should provide more, less, or the same amount of help for African Americans. Among White respondents, it is socially undesirable to express opposition to government programs to help Black Americans (see Holbrook et al. 2003). Hence, we could assess the mode difference in social desirability response bias among Whites.

SAMPLES

OSU Center for Survey Research. Data collection was conducted by a group of 14 supervisors and 59 interviewers; both groups received formal training before working on the project and were continually monitored throughout the field period. Households were selected based on RDD within the 48 contiguous U.S. states, and one adult per household was randomly sampled to be interviewed using the "last birthday" method (Lavrakas 1993). As shown in Table 1, 1,506 respondents were interviewed pre-election between June 1 and July 19, 2000, and 1,206 of those respondents were interviewed post-election between November 9 and December 12, 2000, after the general elections. For the pre-election wave, the AAPOR Response Rate 5 was 43%; the cooperation rate was 51%. Post-election, the number of completions divided by the number of Wave I respondents yielded a response rate of 80%; the cooperation rate was 94%.

Knowledge Networks. AAPOR Contact Rate 2 was about 89% for the initial telephone interview to

recruit people to join the KN panel. 6-8% of households were ineligible because they were outside of the WebTV service area. Of the interviewed eligible respondents, 56% agreed to join the KN panel, and 72% of these people eventually had WebTV installed in their homes.

The pre-election survey was conducted in July 2000, and the post-election survey was conducted in November 2000. The pre-election questionnaire was divided into three separate modules. Respondents were invited to complete the second module one week after they had been invited to complete the first module, and invitations to complete the third module were sent out two weeks after the invitations for the second module were sent out. 7,054 people were invited to complete the first module. 4,933 respondents completed all three modules within four weeks after the invitations for the first module were sent out, yielding a panel completion rate of 70%. Of the 4,933 respondents who completed the entire pre-election questionnaire, 790 were excluded from assignment to the post-election survey for varying reasons.[2] The remaining 4,143 people were invited to complete the post-election survey on November 8, 2000, and 3,416 did so within two weeks after receiving the invitation, yielding an 82% completion rate.

Harris Interactive. In June, 2000, 12,523 participants were pulled from the HPOL database stratified by gender, age, and region of residence (Northeast, South, Midwest, and West). The selected sample matched population parameters (from CPS data) in terms of distributions of age and region of residence, and there was an oversample of male respondents (because HI expected that non-respondents were more likely to be male than female). 2,306 respondents completed the pre-election questionnaire, yielding a completion rate of 18%.

After the election in November, 2000, these respondents were invited to complete the post-election survey, and 1,028 did so, yielding a completion rate of 45%. No incentives were offered to respondents in exchange for their participation in this study.

DEMOGRAPHIC REPRESENTATIVENESS OF THE PRE-ELECTION SAMPLES

The demographics of the American adult population were gauged using the Annual

Demographic Survey supplement of the Current Population Survey (CPS) conducted in March, 2000.[3]

Table 2 displays these data and the demographics of the three pre-election samples. For each house, the left column shows the distributions for the unweighted samples, and the right column shows the distributions for the samples weighted using the weights provided to us by the data collection organizations. Under each column of percentages for a demographic variable is the average deviation of the results from the comparable CPS figures.

Focusing first on the unweighted samples, the CSR sample manifested the smallest average deviation for three variables (education, income, and age), whereas KN manifested the smallest deviations for two other variables (race and gender).[4] The HI sample consistently manifested the largest average deviations from the population. As shown in the bottom row of Table 2, the average deviation for the unweighted samples was 4.0% for CSR, 4.3% for KN, and 8.7% for HI.

Education. The CSR sample under-represented the least educated individuals and over-represented individuals with college degrees or postgraduate degrees. A similar bias was present in the KN sample: people with high school education were under-represented, whereas people with more education were over-represented. The same bias was even more pronounced in the HI sample, which severely under-represented people with some high school education and high school graduates, and substantially over-represented people who had done post-graduate studies.

Income. The CSR sample under-represented the lowest income individuals; this bias was stronger in the KN sample and even more pronounced in the HI sample. All three samples over-represented the highest income individuals

Age. The CSR sample under-represented individuals under age 25 and over age 75, but discrepancies from the population statistics were never large. Discrepancies were larger in the KN sample, which under-represented individuals under age 25 and over age 65. The same biases were most apparent in the HI sample, which substantially under-represented people over age 65.

Race. The CSR sample under-represented African-American respondents, and the KN and HI

samples evidenced this same bias more strongly. The CSR sample under-represented White respondents, whereas the KN and HI samples over-represented Whites. All three samples over-represented people of other races, with the CSR sample doing so the most.

Gender. The CSR sample over-represented women, whereas the HI sample over-represented men. The KN sample's gender composition closely matched the population, and the HI sample was most discrepant.

Impact of Weighting. The CSR weights adjusted for probability of selection using number of voice telephone lines and number of adults in the household, post-stratified using the March 2000 CPS using age, education, income, race, and gender. KN similarly adjusted for unequal probabilities of selection using number of voice telephone lines per household and several sample design features and used rim weighting (with 10 iterations) to adjust according to the most recent monthly CPS figures. The HI weights used CPS data and answers to some questions administered in monthly telephone surveys of national cross-sectional samples of 1,000 adults, aged 18 and older as benchmarks in terms of gender, age, education, race, ethnicity, and a variable representing the propensity of an individual respondent to have regular access to the Internet. In Table 2, the right column under each house's label shows the distributions of the demographics after the weights were applied. Not surprisingly, weighting shrunk the demographic deviations from the population considerably.

DEMOGRAPHIC REPRESENTATIVENESS OF THE POST-ELECTION SAMPLES

Table 3 shows the distributions of the demographics of the post-election samples in the same format as was used in Table 2. Among the unweighted samples, the CSR sample continued to manifest the smallest average deviations for education, income, and age, and the KN sample maintained the smallest deviations for race and gender. The HI sample showed the largest average deviations from the population on all 5 attributes. As shown in the bottom row of the table, the average deviations for the unweighted samples were 4.5% for the CSR sample, 4.3% for the KN sample, and 9.3% for the HI sample. Weighting had a similar effect here to that observed in Table 2.

A number of indicators suggest that the HI respondents were considerably more interested in the topic of the survey (politics) than were the CSR and KN respondents (see Table 4). The CSR respondents gave significantly fewer correct answers to the political knowledge quiz questions than the KN respondents ($\underline{b}$=.09, p<.001 unweighted; $\underline{b}$=.08, p<.001 weighted) and HI respondents ($\underline{b}$=.24, p<.001 unweighted; $\underline{b}$=.19, p<.001 weighted). And the KN respondents gave significantly fewer correct answers than the HI respondents ($\underline{b}$=.16, p<.001 unweighted; $\underline{b}$=.11, p<.001 weighted). The same differences persisted after controlling for sample differences in demographics: the CSR respondents gave significantly fewer correct answers than the KN respondents ($\underline{b}$=.07, p<.001 unweighted; $\underline{b}$=.07, p<.001 weighted) and the HI respondents ($\underline{b}$=.18, p<.001 unweighted; $\underline{b}$=.19, p<.001 weighted), and the KN respondents gave significantly fewer correct answers than the HI respondents ($\underline{b}$=.11, p<.001 unweighted; $\underline{b}$=.12, p<.001 weighted).

Likewise, the rate at which respondents selected the midpoints of rating scales (thereby indicating neutrality in evaluations of politicians, national conditions, and government policies) was highest for the CSR respondents, a bit lower for the KN respondents, and considerably lower for the HI respondents. The CSR respondents manifested significantly more midpoint selection than the KN respondents ($\underline{b}$=-.04, p<.001 unweighted; $\underline{b}$=-.04, p<.001 weighted) and the HI respondents ($\underline{b}$=-.09, p<.001 unweighted; $\underline{b}$=-.10, p<.001 weighted). And the KN respondents manifested significantly more midpoint selection than HI respondents ($\underline{b}$=-.06, p<.001 unweighted; $\underline{b}$=-.06, p<.001 weighted). The same differences persisted after controlling for sample differences in demographics: the CSR respondents manifested significantly more midpoint selection than the KN respondents ($\underline{b}$=-.04, p<.001 unweighted; $\underline{b}$=-.03, p<.001 weighted) and the HI respondents ($\underline{b}$=-.09, p<.001 unweighted; $\underline{b}$=-.09, p<.001 weighted), and the KN respondents manifested significantly more midpoint selection than the HI respondents ($\underline{b}$=-.05, p<.001 unweighted; $\underline{b}$=-.06, p<.001 weighted).

The CSR and KN samples contained comparable proportions of people who identified

themselves as political independents (rather than identifying with a political party), whereas the proportion of independents in the HI sample was considerably lower. The KN and CSR respondents were not significantly different from one another (p>.80 unweighted; p>.20 weighted), whereas the HI respondents were significantly less likely to be independents than the CSR respondents (b=-.58, p<.001 unweighted; b=-.61, p<.001 weighted) or the KN respondents (b=-.59, p<.001 unweighted; b=-.63, p<.001 weighted). The same differences persisted after controlling for sample differences in demographics: a non-significant difference between the KN and CSR respondents (p>.60 unweighted; p>.10 weighted), whereas the HI respondents were significantly less likely to be independents than the CSR respondents (b=-.20, p<.01 unweighted; b=-.22, p<.001 weighted) or the KN respondents (b=-.25, p<.01 unweighted; b=-.40, p<.001 weighted).

The HI respondents were most likely to say pre-election that they intended to vote in the upcoming election, and the KN respondents were least likely to predict they would vote. The CSR respondents were more likely than the KN respondents to say they would vote (b=-.35, p<.001 unweighted; b=-.40, p<.001 weighted) and less likely than the HI respondents to predict they would vote (b=1.06, p<.001 unweighted; b=.58, p<.001 weighted). The KN respondents were less likely than the HI respondents to predict they would vote (b=1.41 p<.001 unweighted; b=.98, p<.001 weighted). The same differences persisted after controlling for sample differences in demographics: the CSR respondents were more likely than the KN respondents to predict they would vote (b=-.60, p<.001 unweighted; b=-.63, p<.001 weighted) and less likely than the HI respondents to predict they would (b=.36, p<.05 unweighted; b=.31, p<.05 weighted). The KN respondents were less likely than the HI respondents to predict they would vote (b=.96, p<.001 unweighted; b=.94, p<.001 weighted).

Post-election reports of voter turnout were about equal in the CSR and KN samples and considerably higher in the HI sample (see the bottom portion of Table 4). The CSR and KN rates were not significantly different from one another unweighted (p>.30), but when the samples were weighted, the CSR respondents' reported turnout rate was significantly higher than that of the KN respondents

($\underline{b}$=-.28, p<.01).[5] The CSR respondents reported significantly lower turnout than the HI respondents, both weighted and unweighted ($\underline{b}$=1.39, p<.001 unweighted; $\underline{b}$=1.21, p<.001 weighted). The KN respondents reported significantly lower turnout than the HI respondents ($\underline{b}$=1.46 p<.001 unweighted; $\underline{b}$=1.35, p<.001 weighted). After controlling for sample differences in demographics, the CSR respondents reported significantly higher turnout than the KN respondents ($\underline{b}$=-.26, p<.01 unweighted; $\underline{b}$=-.38, p<.001 weighted) and significantly lower turnout than the HI respondents ($\underline{b}$=.57, p<.05 unweighted; $\underline{b}$=.56, p<.05 weighted). The KN respondents reported significantly lower turnout than the HI respondents ($\underline{b}$=.83, p<.001 unweighted; $\underline{b}$=.94, p<.001 weighted).

All three samples over-estimated voter turnout as compared to the official figure of 51.3% documented by the Federal Election Commission for the 2000 Presidential elections, the KN sample being closest by a statistically significant margin, and the HI sample being farthest away.

CONCURRENT VALIDITY

Binary logistic regressions were conducted predicting vote choice (coded 1 for Mr. Gore and 0 for Mr. Bush) with a variety of predictors using only respondents who said they expected to vote for Mr. Bush or Mr. Gore.[6] All predictors were coded to range from 0 to 1, with higher numbers implying a more favorable orientation toward Mr. Gore. Therefore, positively signed associations with predicted vote and actual vote were expected.

Concurrent validity varied substantially across the three houses (see Table 5). As shown in the bottom row of Table 5, the average change in probability that a respondent will vote for Gore instead of Bush based on the predictor measures in the CSR sample (unweighted: .47; weighted: .46) was weaker than the average change in probability for KN (unweighted: .56; weighted: .55), which in turn was weaker than the average change in probability for HI (unweighted: .63; weighted: .59). Concurrent validity was significantly lower for CSR than for KN for 22 of the 41 predictors, and concurrent validity was significantly lower for KN than for HI for 34 of the 41 predictors. Concurrent validity was significantly higher for CSR than for KN for none of the 41 predictors, and concurrent validity was

significantly higher for KN than for HI for none of the predictors. Sign tests revealed significantly

lower concurrent validity for CSR than for KN (p<.001), significantly lower concurrent validity for

CSR than for HI (p<.001), and significantly lower concurrent validity for KN than for HI (p<.001).[7]

Some of these differences between houses may be due to differences between the three samples

in terms of demographics and political knowledge. To reassess the house effects after adjusting for

those differences, we concatenated the data from the three houses into a single dataset and estimated

the parameters of regression equations predicting predicted vote choice with each substantive predictor

(e.g., party identification), two dummy variables to represent the three houses, education, income, age,

race, gender, political knowledge, political knowledge squared, and interactions of all of these latter

variables with the substantive predictor. The interactions involving the demographics and knowledge

allowed for the possibility that concurrent validity might vary according to such variables and might

account partly for differences between the houses in observed concurrent validity. Our interest was in

the two interactions of the house dummy variables with the substantive predictor; significant

interactions would indicate reliable differences between houses in concurrent validity.

After controlling for demographics and political knowledge in concatenated regressions, sign

tests again revealed significantly lower predictive validity for CSR than for KN (p<.001), significantly

lower concurrent validity for CSR than for HI (p<.001), and significantly lower concurrent validity for

KN than for HI (p<.001). Applying the sample weights weakened these differences a bit, but sign tests

again revealed significantly lower concurrent validity for CSR than for KN (p<.001) and significantly

lower concurrent validity for KN than for HI (p<.05), even when including the demographics and

political knowledge and their interactions with the predictors in the equations.

PREDICTIVE VALIDITY

Table 6 shows change in probability estimates from equations predicting post-election vote

choice with the 41 potential vote choice determinants. As shown in the bottom row of Table 6, the

average change in probability that a respondent will vote for Gore instead of Bush based on the

predictor measures in the CSR sample (unweighted: .46; weighted: .45) was weaker than the average change in probability for KN (unweighted: .54; weighted: .53), which in turn was weaker than the average change in probability for HI (unweighted: .64; weighted: .57). Predictive validity was significantly lower for CSR than for KN for 24 of the 41 predictors, and predictive validity was significantly lower for KN than for HI for 32 of the 41 predictors. Predictive validity was significantly higher for CSR than for KN for none of the 41 predictors, and predictive validity was significantly higher for KN than for HI for none of the predictors. Sign tests revealed significantly lower predictive validity for CSR than for KN (p<.001), significantly lower predictive validity for CSR than for HI (p<.001), and significantly lower predictive validity for KN than for HI (p<.001).

After controlling for demographics and political knowledge in concatenated regressions, sign tests again revealed significantly lower predictive validity for CSR than for KN (p<.05), significantly lower predictive validity for CSR than for HI (p<.001), and significantly lower predictive validity for KN than for HI (p<.001). Applying the sample weights again weakened these differences, particularly the difference between KN and HI. Sign tests revealed significantly lower predictive validity for CSR than for KN (p<.001), and marginally significantly lower predictive validity emerged for KN than for HI (p<.10).[8]

SURVEY SATISFICING

The CSR respondents manifested more non-differentiation than the KN respondents (unweighted: M=.40 vs. .38, b=-.02, p<.01; weighted: M=.41 vs. .38, b=-.02, p<.001), and the HI respondents manifested the least non-differentiation (unweighted: M=.32, b=-.06, p<.001 compared with KN; weighted: M=.34, b=-.05, p<.001 compared with KN).[9] After controlling for differences between the samples in terms of demographics and political knowledge, the difference between KN and CSR was no longer statistically significant (unweighted p>.20; weighted p>.50), but HI continued to manifest the least non-differentiation (unweighted: b=-.04, p<.001 compared with KN; weighted: b=-.04, p<.001 compared with KN).

RELIABILITY

To gauge the amount of random measurement error in answers using the pre-election and post-election feeling thermometer ratings of Mr. Bush and Mr. Gore, LISREL 8.14 was employed to estimate the parameters of the model shown in Figure 2, which posited a latent candidate preference both pre-election and post-election, measured by the feeling thermometer ratings. The stability of the latent construct is represented by a structural parameter, b_{21}. ε_1 - ε_4 represent measurement error in each indicator, and λ_1 - λ_4 are loadings of the manifest indicators on the latent factors. The larger λ_1 - λ_4 are, the higher the validities of the indicators; the smaller ε_1 - ε_4 are, the higher the reliabilities of the items are.

The parameters of the model were estimated separately for CSR, KN, and HI three times, first unweighted, then weighted using the weights supplied by the survey firms, and finally weighted using a set of weights we built to equate the samples in terms of demographics and political knowledge. Specifically, we weighted each sample to match the age, gender, education, and race benchmarks from the 2000 CPS March Supplement and to match the average political knowledge scores from all three samples combined.[10]

Consistently across all four indicators, the factor loadings were smallest for CSR, intermediate for KN, and largest for HI (see Table 7). The error variances were consistently the largest for CSR, intermediate for KN, and smallest for HI. All of the differences between adjacent columns in Table 7 are statistically significant ($p<.05$). Thus, these results are consistent with the conclusion that the CSR reports were less reliable than the KN reports, which in turn were less reliable than the HI reports.

SOCIAL DESIRABILITY RESPONSE BIAS

Among White respondents, it is socially undesirable to express opposition to government programs to help Black Americans (see Holbrook et al., 2003). When asked whether the federal government should provide more, less, or the same amount of help for African Americans, the distributions of answers from White respondents differed significantly across the three houses. White

KN respondents were more likely than White CSR respondents to say the government should provide less help to Black Americans (unweighted: CSR =17.0% vs. KN = 35.8%, χ^2 = 188.87, p<.001; weighted: CSR = 16.1% vs. KN = 34.1%, χ^2 = 189.41, p<.001). And White HI respondents were more likely than White KN respondents to say the government should provide less help to Black Americans (unweighted: KN = 35.8% vs. HI = 42.5%, χ^2 = 30.98, p<.001; weighted: KN = 34.1% vs. HI = 34.1%, χ^2 = 13.90, p<.001). The same differences persisted when controlling for demographics and political knowledge: White CSR respondents gave significantly fewer socially undesirable answers than White KN respondents (b=.88, p<.001) and White HI respondents (b=1.02, p<.001). And White KN respondents gave significantly fewer socially undesirable answers than White HI respondents (b=.13, p<.05).[11]

We also tested whether these differences persisted when controlling for vote choice in the 2000 Presidential election, party identification, and political ideology. The HI sample was more pro-Republican and more politically conservative than the other samples, so this may have been responsible for the HI sample's greater opposition to government help to Black Americans. And in fact, controlling for these additional variables made the difference in answers to the aid to Blacks question between White KN and HI respondents non-significant (b=.10, p>.10). However, even with these controls, White CSR respondents gave significantly fewer socially undesirable answers than did White KN respondents (b=1.00, p<.001) and White HI respondents (b=1.11, p<.001). Thus, the mode difference persisted.

PAST EXPERIENCE AND SELECTIVITY

The KN and HI data may have manifested higher response quality than the telephone data partly because the Internet respondents were panel members who had more practice doing surveys than the average telephone respondent. So we could test this notion, KN provided the number of invitations sent to each respondent and the number of surveys each respondent completed during the 3 months

prior to our pre-election survey. HI provided the number of invitations sent to each respondent and the number of surveys each respondent ever completed.

We computed two variables: (a) "past experience," number of completed surveys in the past (recoded to range from 0 to 1 in both samples), and (b) "selectivity," the rate of responding to past invitations, which was the number of completions divided by number of invitations (also recoded to range from 0 to 1).[12]

To assess whether past experience or selectivity affected response quality, we repeated the binary logistic regressions predicting vote choice using each of the 41 predictors, controlling for the main effects of past experience and selectivity and the interactions between these two variables with each predictor. If having more experience with surveys improved response quality, a significant positive interaction between past experience and each predictor should appear. If being more selective about survey participation results in higher response quality on the surveys that a person completes, a significant negative interaction between selectivity and each predictor should appear.

These data uncovered many indications that past experience improved survey performance in the KN data. Past experience interacted positively with 37 of 41 predictors in the concurrent validity equations, meaning that concurrent validity was higher for people who had more past experience. Eleven of these interactions were significant (p<.05), and none of the interactions in the opposite direction were significant. In the predictive validity equation, past experience was positively associated with predictive validity for 33 of the 41 predictors in the KN data. Six of these effects were significant, and none of the interactions in the opposite direction were significant.

In contrast, the HI data showed very little evidence of practice effects. Past experience interacted positively with 23 of 41 predictors in the concurrent validity equations, just about the number that would be expected by chance alone. Only three of these interactions were significant, and none of those in the opposite direction were significant. In the predictive validity equations, 24 of the 41 predictors yielded positive interactions, only 3 of which were statistically significant, and none of the

past experience effects in the opposite direction were significant. The absence of practice effects in the HI data may be because the range of practice in that sample was relatively small as compared to the KN sample.

Selectivity in past participation did not appear to be a reliable predictor of response quality in the KN sample. Selectivity interacted negatively with 15 of 41 predictors in the concurrent validity assessments (fewer than would be expected by chance), and none of the interactions was significant. Similarly, selectivity interacted negatively with predictive validity for 12 of the 41 predictors in the KN data, and none of these interactions was significant.

In contrast, selectivity was associated with improved response quality in the HI sample. Selectivity interacted negatively with 33 of 41 predictors in the concurrent validity equations; 15 of these interactions were significant, and none were significant in the opposite direction. In the predictive validity equations, 35 of the 41 predictors manifested negative interactions, 10 of which were significant, and none of the interactions in the opposite direction were significant.

All this suggests that at least some superiority in response quality of the KN sample over the CSR sample may be attributable to practice effects, and some of the superiority in response quality of the HI sample over KN sample may be due to strategic selectivity.

DISCUSSION

These data support a series of conclusions:

(1)   The probability samples were more representative of the nation's population than was the non-probability sample, even after weighting.

(2)   The non-probability sample was biased toward individuals who were highly knowledgeable about and interested in the topic of the survey.

(3)   Self-reports provided via the Internet were more accurate descriptions of the respondents than were self-reports provided via telephone, as manifested by higher concurrent and predictive validity, higher reliability, less satisficing, and less social

desirability bias.

(4)   The practice gained by participants in the KN panel enhanced the accuracy of their

self-reports, but such practice did not enhance the accuracy of reports by members

of the non-probability Internet sample.

(5)   The tendency of non-probability sample members to choose to participate in surveys

on topics of great interest to them made their self-reports more accurate on average

than the self-reports obtained from the less selective KN respondents.

Our findings that practice effects enhance the quality of survey responses (and therefore

advantage probability sample Internet surveys) are in harmony with the large literature in psychology

showing that practice improves performance on complex tasks (e.g., Donovan and Radosevich 1999).

And our findings are in line with other evidence suggesting that survey respondents provide more

accurate reports after gaining practice by completing questionnaires (e.g., Novotny, Rumpler, Judd,

Riddick, Rhodes, McDowell, and Briefel 2001).

Although the response rate for the KN sample (25%) was considerably lower than the response

rate for the CSR sample (43%), the average demographic representativeness of the KN sample was

equal to that of the CSR sample. This evidence is consistent with past findings suggesting that declines

in response rates were not associated with notable declines in sample representativeness (Curtin,

Presser, and Singer 2000; Keeter, Miller, Kohut, Groves, and Presser 2000).

A LABORATORY EXPERIMENT

To ascertain whether the differences observed in the national field experiment were merely due

to sample differences, we conducted a controlled laboratory experiment in which respondents were

randomly assigned to provide data either via an intercom that simulated telephone interviews, or via

computers that simulated self-administered web surveys. All respondents answered the same questions,

which were modeled after those used in the national field experiment. The method and results of this

laboratory experiment are described in Appendix 3. In essence, data collected via computers manifested

higher concurrent validity than data collected via intercom. In addition, we found more satisficing in the intercom data than the computer data, as evidenced by more non-differentiation and stronger response order effects. This evidence suggests that features of the computer mode may have facilitated optimal responding.

Replicating results from the national field experiment, computer respondents in the lab experiment were apparently more willing to provide honest answers that were not socially admirable. This finding is consistent with other evidence that eliminating interaction with an interviewer increases willingness to report opinions or behaviors that are not respectable (Sudman and Bradburn 1974; Tourangeau and Smith 1996; Wiseman 1972; Wright et al. 1998).

Conclusion

Taken together, the results from the national field experiment and the laboratory experiment suggest that the Internet offers a viable means of survey data collection and has advantages over telephone interviewing in terms of response quality. These results also demonstrate that probability samples yield more representative results than do non-probability samples. We look forward to future studies comparing data quality across these modes to complement the evidence reported here and to assess the generalizability of our findings.

Endnotes

[1] The 2003 Respondent Cooperation and Industry Image Survey conducted by the Council for Marketing and Opinion Research (CMOR) suggested that 51% of their respondents had participated in surveys within the past year, an average of 5 times (Miller and Haas, 2003).

[2] These included people who were temporarily on inactive status (e.g., on vacation, experiencing health problems, or too busy), people who had been dropped from the panel, and people who were assigned to complete other surveys instead.

[3] The CPS is a monthly survey administered by the Census Bureau using a sample of some 50,000 households. Selected households participate in the CPS for 4 consecutive months, take 8 months off, and then return for another 4 months before leaving the sample permanently. Participants in the CPS are 15 years old or older and are not institutionalized nor serving in the military. The questionnaire is administered via either telephone or face-to-face interviewing.

[4] The initial sample of panel members invited to do the pre-election KN survey was very similar to the subset of those individuals who completed the survey, so discrepancies of the KN sample from the population were largely due to unrepresentativeness of the sample of invited people, rather than due to biased attrition among these individuals who declined to complete the questionnaire.

[5] This result can be viewed as consistent with evidence to be reported later that telephone respondents are more likely than Internet respondents to distort their reports of attitudes and behavior in socially desirable directions.

[6] 26.8% of the CSR respondents, 27.3% of the KN respondents, and 13.5% of the HI respondents predicted that they would vote for someone other than Mr. Bush or Mr. Gore or said they would not predict for whom they would vote despite the follow-up leaning question. All regressions were ran in STATA, which provides correct variance estimates from weighted analyses.

[7] This sign test was computed by assigning a "+" to a predictor if one house had a stronger coefficient than the other and a "-" is assigned if the reverse was true and then computing the probability that the observed distributions of plusses and minuses occurred by chance alone.

[8] For both the pre-election and post-election surveys, the HI sample weights had an unconventionally wide range of values (from 0 to 26). As a result, variance estimates obtained from the weighted HI data were often much larger than those obtained from the other two samples, hence handicapping the ability to detect statistical significance of differences between HI data and the other two houses. The distribution of HI weights was examined for skewness and clumps. Although huge weights were assigned to some respondents, the majority of respondents received weights within the conventional range of less than 3. Furthermore, a sensitivity analysis on change in estimates before and after truncating the weights revealed little change in point estimates and variance estimates in the vote choice regression models presented in this paper. This is not surprising because the huge weights were assigned to very few respondents.

[9] Our method for calculating non-differentiation is explained in Appendix 2.

[10] This weighting was also done using income as well, and the results were comparable to those described in the text.

[11] These logistic regressions predicted socially undesirable responding (coded 1 = "less help for Black Americans" and 0= "same" or "more help for Black Americans") with 2 dummy variables representing the three survey firms and main effects of education, income, age, gender, race, political knowledge, and political knowledge squared.

[12] 10% of HI respondents had never completed any HI survey before the pre-election survey in the present study, whereas only 0.3% of KN respondents had never completed any KN survey prior to ours. So the KN respondents were a bit more experienced with the survey platform than were the HI respondents. About 54% of KN respondents had completed all the surveys that KN had invited them to do during the prior three months, whereas only 2% of the HI respondents had a perfect completion

rate since joining the HPOL panel. Thus, the HI respondents were apparently more selective than were the KN respondents, who were obligated to complete all surveys in order to keep their free WebTV equipment.

[13] Respondents also reported their opinions on seven other policy issues, but the associations between opinions on these issues and vote choices were either zero or close to zero (logistic regression coefficients of .29 or less when the three samples were combined). Therefore, we focused our analyses on the issues that manifested substantial concurrent and predictive validity (logistic regression coefficients of 1.00 or more when the three samples were combined).

[14] Policy preferences on pollution by businesses did not predict the difference in feeling thermometer ratings regardless of mode and were therefore excluded from our concurrent validity analyses.

[15] For efficiency, the massive tables showing detailed coefficients for all main effects and interaction effects are not presented here. These tables are available from the authors upon request.

Appendix 1

Question wordings were closely modeled on items traditionally included in the National Election Studies questionnaires.

Predicted Voter Turnout and Candidate Preference. Pre-election, all respondents were asked whether they expected to vote in the presidential election. Respondents who said they probably would vote were then asked to predict for whom they would probably vote. In all surveys, the names of George W. Bush and Al Gore were listed, and other names were accepted as answers. Respondents who said they were not sure or were undecided were asked to make their best prediction nonetheless.

Feeling Thermometer Ratings. Respondents rated how favorable or unfavorable they felt toward three politicians on a feeling thermometer ranging from 0 to 100: Bill Clinton, Mr. Gore, and Mr. Bush. Larger numbers indicated more favorable evaluations. The midpoint of 50 was labeled as indicating that the respondent felt neither favorable nor unfavorable. The order of presentation of the three names was rotated randomly across respondents.

Approval of President Clinton's Job Performance. Respondents reported their approval of President Clinton's handling of his job as president overall, as well as his handling of the U.S. economy, U.S. relations with foreign countries, crime in America, relations between Black Americans and White Americans, and pollution and the environment (our five "target performance issues"). Ratings were provided on a 5-point scale ranging from "strongly approve" to "strongly disapprove."

Perceived Changes in Past National Conditions. Respondents reported their perceptions of whether national conditions on the five target performance issues were currently better than, worse than, or the same as they had been 8 years before, when President Clinton took office. Ratings were provided on a 5-point scale ranging from "much better" to "much worse."

Expectations of National Conditions if Each Candidate Were to Be Elected. Respondents reported

whether they thought national conditions on the five target performance issues would become better or worse during the next 4 years if either Mr. Gore or Mr. Bush were to be elected president. The order of the candidates was rotated randomly across respondents, so that the questions about Mr. Gore appeared before the questions about Mr. Bush for about half the respondents. Ratings were provided on a 5-point scale ranging from "much better" to "much worse." To yield comparative ratings, expectations for conditions if Mr. Bush were to be elected were subtracted from expectations for conditions if Mr. Gore were to be elected.

Perceptions of Candidates' Personality Traits. Respondents reported the extent to which four personality trait terms described Mr. Gore and Mr. Bush: moral, really cares about people like you, intelligent, and can provide strong leadership. Again, the questions about the two candidates were randomly rotated across respondents, as was the order of the four traits. Ratings were made on 4-point scale ranging from "extremely" to "not at all." Ratings of Mr. Bush on each trait dimension were subtracted from ratings of Mr. Gore on that dimension to yield comparative scores for each trait.

Emotions Evoked by the Candidates. Respondents reported the extent to which Mr. Gore and Mr. Bush made them feel each of four emotions: angry, hopeful, afraid, and proud. The order of the questions about each candidate was randomly rotated across respondents, as was the order of the emotions. Ratings were made on a 5-point scale ranging from "extremely" to "not at all." Ratings of Mr. Bush were subtracted from ratings of Mr. Gore to yield comparative scores on each emotion.

Policy Preferences. Respondents were asked what they thought the government should do on a number of policy issues, using two question formats. The first format asked respondents to report whether they thought there should be increases or decreases in government spending on the military, government spending on social welfare programs, government help for African Americans, the strictness of gun control, the strictness of regulations limiting environmental pollution by businesses, efforts to fight crime, and restrictions on immigration. Ratings were made on a 5-point scale ranging from "a lot less" to "a lot more."

Respondents were also asked whether it was or would be a good or bad thing for the government to pursue certain policy goals, including: making abortion illegal under all circumstances, making abortion legal under all circumstances, helping poor countries provide food, clothing, and housing for their people, preventing people in other countries from killing each other, preventing governments in other countries from killing their own citizens, helping to resolve conflicts between two other countries, preventing other countries from polluting the environment, and building weapons to blow up missiles fired at the US.[13] These ratings were made on 5-point scales ranging from "very good," "somewhat good," "neither good nor bad," "somewhat bad," to "very bad."

Party Identification. Party identification was measured by asking respondents: "Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent, or what?" Respondents who chose either Republican or Democrat were then asked, "Would you call yourself a strong [Republican/Democrat] or a not very strong [Republican/Democrat]?" Respondents who said they were Independents were asked instead, "Do you think of yourself as closer to the Republican Party, closer to the Democratic Party, or equally close to both?" Responses were used to build a 7-point party identification scale. The CSR and HI respondents answered these questions pre-election; for KN respondents, these questions were in one of the profile surveys they completed when they joined the panel.

Political Ideology. Respondents indicated their political ideology by selecting one of five response options: very liberal, liberal, moderate, conservative, or very conservative. CSR and HI respondents provided their ratings pre-election; and KN respondents answered these questions during one of their initial profile surveys.

Political Knowledge. Five questions measured respondents' factual political knowledge (Delli Carpini and Keeter 1996; Zaller 1992): (a) Do you happen to know what job or political office is now held by Trent Lott? (b) Whose responsibility is it to determine if a law is constitutional or not? (c) How much of a majority is required for the U.S. Senate and House to override a presidential veto? (d) Which

political party currently has the most members in the House of Representatives in Washington? (e) Would you say that one of the political parties is more conservative than the other? (If yes) Which party would you say is more conservative? Correct answers were "Senator majority leader," "Republican Senator," "Senator," or "Senator from Mississippi" for question (a), "Supreme Court" for question (b), "two-thirds" for question (c), and "Republicans" for questions (d) and (e). Each respondent was given a composite score: the percent of correct answers given, ranging from 0 to 1.

Demographics. Age was computed from respondents' reported year of birth. Respondents reported their race in five categories: White, Black, Native American, Asian/Pacific Islander, and other. The gender of respondents was noted by the telephone interviewers and was reported by the Internet respondents. Respondents indicated the highest level of education they had completed, with the list of response options ranging from "less than high school" to "completed graduate school."

Internet respondents were given a list of income categories for reporting their 1999 household income. The range of the 17 KN categories was from "less than $5,000" to "more than $125,000," whereas the range of the 16 HI categories was from "less than $10,000" to "more than $250,000." Telephone respondents were first asked to state a figure for their household income in 1999. Respondents who did not give a number were read a series of nine categories, one at a time, and were asked to stop the interviewer when their income category was reached. These categories ranged from "less than $10,000" to "more than $150,000."

Mode Differences in Presentation Format. Almost all questions and response scales were identical in the two modes, but a few adaptations were made in wording and formatting to suit each mode. For example, telephone interviewers used the pronoun "I" to refer to themselves, whereas the pronoun "we" was used on the Internet surveys to refer to the researchers. The telephone survey presented some questions in branching formats, whereas the Internet survey presented those questions in grid formats without branching. All response options were arrayed across the tops of the grids, and multiple question stems were listed down the left side (e.g., different aspects of President Clinton's job

performance were listed together on a grid). This formatting difference represents a typical distinction between the ways rating scale questions are presented to respondents in these modes. Previous research has shown that branching typically yields more reliable and valid judgments than a non-branching format (Krosnick and Berent 1993), so this difference probably advantaged the CSR data in terms of response quality.

POST-ELECTION MEASURES

Turnout. Respondents were asked whether they usually voted in past elections and whether they voted in the 2000 Presidential election.

Vote Choice. Respondents who said they voted in the 2000 Presidential Election were then asked: "Who did you vote for in the election for President, Al Gore, the Democrat, George W. Bush, the Republican, or someone else?"

Feeling Thermometer Ratings. Respondents rated Mr. Clinton, Mr. Gore, and Mr. Bush on the 101-point feeling thermometer.

Appendix 2

To compute the non-differentiation score for each respondent, we used the three pre-election feeling thermometer ratings and a formula developed by Mulligan, Krosnick, Smith, Green, and Bizer (2001):

$$\overline{x_1} = \left(\frac{\sqrt{|therm1 - therm2|} + \sqrt{|therm1 - therm3|} + \sqrt{|therm2 - therm3|}}{3} \right)$$

Because thermometer ratings had been recoded to range from 0 to 1, scores on this index ranged from 0 to .804. A score of 0 indicated that all three thermometer ratings were identical, and a score of .804 indicated the highest level of observed differentiation among thermometer ratings. To yield an index where higher scores indicated more non-differentiation, we subtracted .804 from each score and divided it by –.804, yielding a non-differentiation index that ranged from 0 (indicating the least non-differentiation) to 1 (indicating the most differentiation).

Appendix 3

The national field experiment showed that data collected from national samples via the Internet manifested higher concurrent and predictive validity, higher reliability, less systematic measurement error, and less social desirability response bias than did data collected via national RDD telephone interviewing. However, it is impossible to tell from a field experiment how much of the apparent difference between modes in data quality can be attributed to sample differences among the three houses.

To assess whether the differences observed in the national field experiment were due to sample differences, we conducted a controlled laboratory experiment in which respondents were randomly assigned to provide data either via simulated telephone interviews or via computers simulating self-administered web surveys. All respondents answered the same questions, which were modeled after those used in the national surveys. We compared the concurrent validity of responses, the extent of satisficing, and the extent of social desirability response bias in the two modes.

METHODOLOGY

Respondents. Respondents were undergraduates enrolled in introductory psychology classes at Ohio State University during Spring 2001. They accessed an online database of all experiments available for participation that quarter and chose to sign up for this experiment in exchange for course credit. Only people who had resided in the United States for at least the past 5 years were eligible to participate. The respondents included 174 males and 158 females, most of them born between 1979 and 1982; 78% of the respondents were White, 11% were African-American, 2% were Hispanic, 6% were Asian, and the remaining 3% were of other ethnicities.

Procedure. Respondents arrived at the experimental lab at scheduled times in groups of 4-6 and were each individually randomly assigned to soundproof cubicles. Each cubicle contained either a computer on which to complete a self-administered questionnaire or intercom equipment. Respondents completed the questionnaire by their assigned mode and were debriefed and dismissed.

Interviewers. The interviewers were experienced research assistants who received training on how to administer the questionnaire, record answers, and manage the interview process. The procedures used for training these interviewers were those used by the Ohio State University Center for Survey Research. Following training, the interviewers practiced administering the questionnaire on the intercom. They were closely monitored during the interviewing process, and regular feedback was provided, as would be standard in any high-quality survey data collection firm.

MEASURES

The questions included many items similar to those used in the national surveys, with eight people on a 101-point thermometer scale: Bill Clinton, Al Gore, George W. Bush, Dick Cheney, Colin Powell, Jesse Jackson, Janet Reno, and John Ashcroft. Approval of President Clinton's job p performance were rated on 7 issues: the U.S. economy, U.S. relations with foreign countries, crime in America, education in America, relations between Black Americans and White Americans, pollution and the environment, and health care in America, as were perceived changes in past national conditions. Respondents judged whether national conditions on these 7 target performance issues would become better or worse during the next 4 years under two scenarios: (1) given that George W. Bush was elected president, and (2) if instead Al Gore had been elected. They also rated the extent to which four personality traits described each of the two presidential candidates: moral, really cares about people like you, intelligent, can provide strong leadership. Measures of emotions evoked by the candidates were again: angry, hopeful, afraid, and proud. Policy preferences were tapped in the areas of military spending, social welfare spending, help for African Americans, the strictness of gun control laws, regulation of environmental pollution by businesses, effort to fight crime, and restrictions on immigration. Ratings were made on 5-point scales ranging from "a lot more" to "a lot less," with a midpoint of "about the same." Responses were coded to range from 0 to 1, with larger numbers meaning a disposition more likely to favor Mr. Gore (less military spending, more restriction on pollution by businesses, less immigration restrictions, more welfare

spending, more help for African Americans, stricter gun control, and more efforts to control crime).

Respondents also indicated their party identification and ideology.

In addition, other items not present in the national survey were included in the experiment.

Respondents were asked to identify the most important problem facing the country, the most

important problem facing young people in the country, the most important environmental problem

facing the country, and the most important international problem facing the country. Each question

offered respondents 4 response options. Half of the respondents (selected randomly) were offered

the options in sequence A, B, C, D, whereas the other half were offered the options in sequence D,

C, B, A. In addition, 205 of the 332 respondents granted permission authorizing us to obtain their

verbal and math SAT or ACT test scores from the University Registrar's office. All ACT scores were

converted into SAT scores using the concordance table available at the College Board website

(www.collegeboard.com), showing the equivalent SAT scores for each corresponding ACT score.

Total SAT scores were recoded to range from 0 to 1; the lowest total score of 780 was coded 0, and

the highest total score of 1480 was coded 1.

CONCURRENT VALIDITY

Table 8 displays unstandardized regression coefficients estimating the effects of 38

postulated predictors on the feeling thermometer ratings of Mr. Bush subtracted from feeling

thermometer ratings of Mr. Gore.[14] The computer data yielded significantly higher concurrent

validity than the intercom data for 29 of these predictors. In no instance did the intercom data

manifest significantly higher concurrent validity than the computer data. Across all coefficients

shown in Table 8, Sign tests revealed statistically significantly higher concurrent validity in the

computer data than in the intercom data ($p < .001$).

To explore whether the mode difference varied in magnitude depending upon individual

differences in cognitive skills, we regressed the difference in thermometer ratings on each predictor,

a dummy variable representing mode, cognitive skills, and two-way interactions of mode x the

predictor, cognitive skills x the predictor, and mode x cognitive skills, and the three-way interaction of mode x the predictor x cognitive skills.[15] The three-way interaction tested whether the mode effect on concurrent validity was different for people with varying levels of cognitive skills. We estimated the parameters of this equation using each of the 38 predictors listed in Table 8.

The three-way interaction was negative for 84% (32) of the predictors (7 of them statistically significant) and positive for 6 predictors (none statistically significant). Sign tests revealed that the three-way interaction was more likely to be negative than positive ($p<.001$), indicating that the mode difference was more pronounced among respondents with limited cognitive skills. Among participants in the bottom quartile of cognitive skills (N=52), the computer data yielded significantly higher concurrent validity than the intercom data for 16 out of 38 predictors; whereas among participants in the top quartile of cognitive skills (N=53), the two modes did not yield statistically significantly different concurrent validity for any of the 38 predictors. Thus, it seems that respondents high in cognitive skills could manage the two modes equally well, whereas respondents with more limited cognitive skills were especially challenged by oral presentation.

SURVEY SATISFICING

Non-differentiation. Non-differentiation was measured using responses to the eight feeling thermometer questions with a formula developed by Mulligan *et al.* (2001). Values can range from 0 (meaning the least non-differentiation possible) to 1 (meaning the most non-differentiation possible). Intercom respondents (M=.50) manifested significantly more non-differentiation than the computer respondents on the feeling thermometers (M=.44), $\underline{t}$=3.14, $\underline{p}$<.01. To test whether the mode difference in satisficing was contingent on individual differences in cognitive skills, we ran an OLS regression predicting the non-differentiation index using mode, cognitive skills, and the interaction between mode and cognitive skills. The interaction was negative and statistically significant, indicating that the mode difference in non-differentiation was more pronounced among respondents with more limited cognitive skills (b=-.15, p<.05).

Response Order Effects. When asked the four "most important problem" questions, half of the respondents were offered the response options in the order of A, B, C, D, whereas the other half were offered the options in the order of D, C, B, A. We computed a composite dependent variable by counting the number of times each respondent picked response option A or B, which were the first or second response option for half of the respondents, and the third or fourth response option for the other half. This composite variable ranged from 0 to 4, where 0 indicates that a respondent never picked response option A or B across all four "most important problem" items, and 4 indicates that a respondent always picked response option A or B. Then, within each mode, this composite dependent variable was regressed on a dummy variable representing response choice order (coded 0 for people given order A, B, C, D and 1 for people given order D, C, B, A).

A significant recency effect emerged in the intercom mode (b=.49, p<.01), indicating that response choices were more likely to be selected if they were presented later than if they were presented earlier. In contrast, no response order effect was evident in the computer mode (b=.07, p>.60). When the composite dependent variable was regressed on the dummy variable representing response choice order, cognitive skills, and the 2-way interaction between response choice order and cognitive skills, a marginally significant interaction effect emerged among respondents in the intercom mode (b=1.77, p<.10). This interaction indicates that the mode difference was substantial among people with stronger cognitive skills (computer: b=-.10, ns., N=57; intercom: b=.68, p<.05, N=68) and invisible among respondents with more limited cognitive skills (computer: b=.17, ns., N=49; intercom: b=.21, ns., N=49).

SOCIAL DESIRABILITY RESPONSE BIAS

As in the national field experiment, we explored whether social desirability response bias varied across the modes using the question asking whether the federal government should provide more or less help for African Americans. The distributions of answers from White respondents differed significantly across the two modes, χ^2 = 16.78, p<.01. White intercom respondents were

more likely than White computer respondents to say the government should provide more help to Black Americans (49% in intercom mode vs. 36% in computer mode), whereas White computer respondents were more likely to say the government should provide less help to Black Americans (16% in intercom mode vs. 38% in computer mode). This suggests that the computer respondents were more comfortable offering socially undesirable answers than were the intercom respondents.

COMPLETION TIME

One possible reason why the intercom interviews might have yielded lower response quality is the pace at which they were completed. If the lack of visual contact in intercom interactions leads interviewers and respondents to avoid awkward pauses and rush through the exchange of questions and answers, whereas self-administration allows respondents to proceed at a more leisurely pace, then the completion times for the intercom interviews may have been less than the completion times for the computer questionnaire completion.

In fact, however, the intercom interviews took significantly longer to complete than the self-administered surveys on computers, $t(330) = 21.68$, $p < .001$. Respondents took an average of 17.3 minutes to complete the self-administered questionnaire, whereas the intercom interviews lasted 26.6 minutes on average.

References

Acree, Michael, Maria Ekstrand, Thomas J. Coates, and Ron Stall. 1999. "Mode Effects in Surveys of Gay Men: A Within-Individual Comparison of Responses by Mail and by Telephone." *Journal of Sex Research* 36:67-75.

Aneshensel, Carol, Ralph Frerichs, Virginia Clark, and Patricia Yokopenic. 1982. "Measuring Depression in the Community: A Comparison of Telephone and Personal Interviews." *Public Opinion Quarterly* 46:111-121.

Aronson, Elliot, Phoebe C. Ellsworth, J. Merrill Carlsmith, and Marti Hope Gonzales. 1990. *Methods of Research in Social Psychology.* New York: McGraw-Hill Publishing Company.

Berrens, Robert P., Alok K. Bohara, Hank Jenkins-Smith, Carol Silva, and David L. Weimer. 2003. "The Advent of Internet Surveys for Political Research: A Comparison of Telephone and Internet Samples." *Political Analysis* 11:1-22.

Best, Samuel J., Brian Krueger, Clark Hubbard, and Andrew Smith. 2001. "An Assessment of the Generalizability of Internet Surveys." *Social Science Computer Review* 19:131-145.

Biemer, Paul. 2001. "Nonresponse Bias And Measurement Bias in a Comparison of Face to Face and Telephone Interviewing." *Journal of Official Statistics* 17:295-320.

Bracht, G. H. and Glass, G. V. 1968. "The External Validity of Experiments" *American Educational Research Journal* 5:437-474.

Bridge, R. Gary, Leo G. Reeder, David Kanouse, Donald R. Kinder, Vivian T. Nagy, and Charles Judd. 1977. "Interviewing Changes Attitudes – Sometimes." *Public Opinion Quarterly* 41:57-64.

Cannell, Charles F., Peter V. Miller and Lois Oksenberg. 1981. "Research on Interviewing Techniques." *Sociological Methodology* 12:389-437.

Chang, LinChiat. 2001. *A Comparison Of Samples And Response Quality Obtained From RDD Telephone Survey Methodology and Internet Survey Methodology.* Doctoral Dissertation, Ohio State University, Columbus, OH.

Chartrand, Tanya L., and John A. Bargh. 1999. "The Chameleon Effect: The Perception-behavior Link and Social Interaction." *Journal of Personality and Social Psychology* 76:893-910.

Clausen, Aage R. 1968. "Response Validity: Vote Report." *Public Opinion Quarterly* 32:588-606.

Clinton, Joshua D. 2001. *Panel Bias from Attrition and Conditioning: A Case Study of the Knowledge Networks Panel.* Stanford.

Cordell, Warren N. and Henry A. Rahmel. 1962. "Are Nielsen Ratings Affected by Non-cooperation, Conditioning, or Response Error?" *Journal of Advertising Research* 2:45-49.

Couper, Mick P. 2000. "Web Surveys: A Review of Issues and Approaches." *Public Opinion Quarterly* 64:464-494.

Curtin, Richard, Stanley Presser, and Eleanor Singer. 2000. "The Effects of Response Rate Changes on the Index of Consumer Sentiment." *Public Opinion Quarterly* 64:413-428.

de Leeuw, Edith D. 1992. *Data Quality in Mail, Telephone and Face to Face Surveys.* Amsterdam: T.T.-publikaties.

de Leeuw, Edith D., and Martin Collins. 1997. "Data Collection Methods and Survey Quality: An Overview." In *Survey Measurement and Process Quality*, ed. Lars E. Lyberg, Paul Biemer, Martin Collins, Edith de Leeuw, Cathryn Dippo, Norbert Schwarz, and Dennis Trewin. New York: John Wiley and Sons.

de Leeuw, Edith, Gideon Mellenbergh, and Joop Hox. 1996. "The Influence of Data Collection Method on Structural Models: A Comparison of a Mail, a Telephone, and a Face-To-Face Survey." *Sociological Methods & Research* 24: 443-472.

de Leeuw, Edith and J. van der Zouwen. 1988. "Data Quality in Telephone and Face To Face Surveys: A Comparative Meta-Analysis." In *Telephone survey methodology*, eds. Robert M. Groves, Paul P. Biemer, Lars E. Lyberg, J. T. Massey, William L. Nicholls II, Joseph Waksberg. New York: John Wiley and Sons, Inc.

Dillman, Don A. 1978. *Mail and Telephone Surveys: The Total Design Method.* New York: John Wiley & Sons.

Dillman, Don A. 2000. *Mail and Internet Surveys: The Tailored Design Method.* New York: Wiley-Interscience.

Donovan, John J., and David J. Radosevich. 1999. "A Meta-Analytic Review of the Distribution of Practice Effect: Now You See It, Now You Don't." *Journal of Applied Psychology* 84:795-805.

Falaris, Evangelos M. and H. E. Peters. 1998. " Survey Attrition and Schooling Choices." *The Journal of Human Resources* 33(2):531-54.

Fitzgerald, John, Peter Gottschalk, and Robert Moffitt. 1998a. "An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics." NBER Technical Working Papers National Bureau of Economic Research, Inc.

Fitzgerald, John, Peter Gottschalk, and Robert Moffitt. 1998b. "An Analysis of the Impact of Sample Attrition on the Second Generation of Respondents in the Michigan Panel Study of Income Dynamics." *The Journal of Human Resources* 33(2):300-344.

Fitzsimons, Gavan J. and Vicki Morwitz. 1996. "The Effect of Measuring Intent on Brand Level Purchase Behavior." *Journal of Consumer Research,* 23:1-11.

Flemming, Greg, and Sonner, Molly. 1999. "Can Internet Polling Work? Strategies for Conducting Public Opinion Surveys Online." Paper presented at the annual meeting of the American

Association for Public Opinion Research, St. Petersburg Beach.

Fournier, Louise, and Vivianne Kovess. 1993. "A Comparison of Mail and Telephone Interview Strategies for Mental Health Surveys." *The Canadian Journal of Psychiatry* 38: 525-535.

Fowler, Floyd Jackson, Anthony M. Roman, and Zhu Xiao Di. 1998. "Mode Effects in a Survey of Medicare Prostate Surgery Patients." *Public Opinion Quarterly* 62:29-46.

Fricker, Scott, Mirta Galesic, Roger Tourangeau and Ting Yan. 2005. "An Experimental Comparison of Web and Telephone Surveys." *Public Opinion Quarterly* 69:370-392.

Gano-Phillips, Susan and Frank D. Fincham. 1992. "Assessing Marriage via Telephone Interviews and Written Questionnaires: A Methodological Note." *Journal of Marriage and the Family* 54:630-635.

Granberg, Donald, and Soren Holmberg. 1991. "Self-reported Turnout and Voter Validation." *American Journal of Political Science* 35:448-459.

Greenwald, Anthony G, Catherine G. Carnot, Rebecca Beach and Barbara Young. 1987. "Increasing Voting Behavior by Asking People if They Expect to Vote." *Journal of Applied Psychology* 72:315-318.

Groves, Robert. 1978. "On the Mode of Administering a Questionnaire and Responses to Open-Ended Items." *Social Science Research* 7: 257-271.

Groves, Robert M., and Robert L. Kahn. 1979. *Surveys by Telephone: A National Comparison with Personal Interviews.* New York: Academic Press.

Groves, R. M., E. Singer, and A. D. Corning. 2000. "Leverage-Saliency Theory of Survey Participation: Description and an Illustration." *Public Opinion Quarterly* 64:299-308.

Herzog, A. Regula, and Willard Rodgers. 1988. "Interviewing Older Adults: Mode Comparison Using Data from a Face-to-face survey and a Telephone Survey." Public Opinion Quarterly 52: 84-99.

Himmelfarb, Samuel and Fran H. Norris. 1987. "An Examination of Testing Effects in a Panel Study of Older Persons." *Personality and Social Psychology Bulletin* 13:188-209.

Holbrook, Allyson, L., Melanie C. Green, and Jon A. Krosnick. 2003. "Telephone vs. Face-to-face Interviewing of National Probability Samples with Long Questionnaires: Comparisons of respondent satisficing and social desirability response bias." *Public Opinion Quarterly* 67: 79-125.

Holbrook, A. L., Krosnick, J. A., & Pfent, A. M. (2007). Response rates in surveys by the news media and government contractor survey research firms. In J. Lepkowski, B. Harris-Kojetin, P. J. Lavrakas, C. Tucker, E. de Leeuw, M. Link, M. Brick, L. Japec, & R. Sangster (Eds.), Telephone survey methodology. New York: Wiley.

Hox, Joop, and Edith de Leeuw. 1994. "A Comparison of Nonresponse in Mail, Telephone, and Face-to-face surveys." *Quality & Quantity* 28: 329-344.

Jagodzinski, Wolfgang, Steffen M. Kuhnel, and Peter Schmidt. 1987. "Is There a "Socratic Effect" in Nonexperimental Panel Studies? Consistency of an Attitude toward Guestworkers." *Sociological Methods and Research* 15:259-302.

Keeter, Scott, Carolyn Miller, Andrew Kohut, Robert M. Goves, and Stanley Presser. 2000. "Consequences of Reducing Nonresponse in a National Telephone Survey." *Public Opinion Quarterly* 64:125-148.

Kenny, David A. 1979. *Correlation and Causality.* New York : Wiley.

Kiecker, Pamela, and James E. Nelson. 1996. "Do Interviewers Follow Telephone Survey Instructions?" *Journal of the Market Research Society* 38:161-176.

Krosnick, Jon A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5:213-236.

Krosnick, Jon A. 1999. "Survey Methodology." *Annual Review of Psychology* 50:537-567.

Lavrakas, Paul J. 1993. *Telephone survey methods: Sampling, selection, and supervision.* Thousand Oaks: Sage Publications.

Lavrakas, Paul J. 1997. "Methods for Sampling and Interviewing in Telephone Surveys." In *Handbook of Applied Social Research Methods*, ed. Leonard Bickman and Debra J. Rog. Thousand Oaks: Sage Publications.

Leary, Mark R. 1995. *Behavioral Research Methods.* Pacific Grove: Brooks/Cole Publishing Company.

Lerner, Jennifer S., Julie H. Goldberg, and Philip E. Tetlock. 1998. "Sober Second Thought: The Effects of Accountability, Anger, and Authoritarianism on Attributions of Responsibility." *Personality and Social Psychology Bulletin* 24: 563-574.

Lerner, Jennifer S., and Philip E. Tetlock. 1999. "Accounting for the Effects of Accountability." *Psychological Bulletin* 125:255-275.

Lubin, B., E. E. Levitt, and M. Zuckerman. 1962. "Some Personality Differences Between Responders and Nonresponders to a Survey Questionnaire." *Journal of Consulting Psychology* 26(2):192.

Lyberg, Lars, and Daniel Kasprzyk. 1991. "Data Collection Methods and Measurement Error: An Overview." In *Measurement Errors in Surveys*, ed. Paul Biemer, Robert M. Groves, Lars E. Lyberg, Nancy Mathiowetz, and Seymour Sudman. New York: John Wiley and Sons.

Mann, Christopher B. 2005. "Unintentional Voter Mobilization:Does Participation in Pre-election Surveys Increase Voter Turnout?" *Annals of the American Academy of Political and Social Science* 601(1):155-168.

Martin, Jean, Colm O'Muircheartaigh, and J. Curtis. 1993. "The use of CAPI for attitude surveys: An experimental comparison with traditional methods." *Journal of Official Statistics* 9:641-661.

Menard, Scott. 1991. *Longitudinal Research*. Newbury Park: Sage Publications.

Neumann, Roland, and Fritz Strack. 2000. "Mood Contagion: The Automatic Transfer of Mood Between Persons." *Journal of Personality and Social Psychology* 79: 211-223.

Novotny, Janet A., William V. Rumpler, Joseph T. Judd, Howard Riddick, Donna Rhodes, Margaret McDowell, and Ronette Briefel. 2001. "Diet Interviews of Subject Pairs: How Different Persons Recall Eating the Same Foods." *Journal of the American Dietetic Association* 101:1189-1193.

Price, Kenneth H. 1987. "Decision Responsibility, Task Responsibility, Identifiability, and Social Loafing." *Organizational Behavior and Human Decision Processes* 40: 330-345.

Rockwood, Todd, Roberta Sangster, and Don A. Dillman. 1997. "The Effect of Response Categories on Questionnaire Answers Context and Mode Effects" *Sociological Methods & Research* 26:118-140

Rohde, Gregory L., and Robert Shapiro. 2000. *Falling Through the Net: Toward Digital Inclusion*. U.S. Department of Commerce: Economics and Statistics Administration and National Telecommunications and Information Administration.

Rossi, Peter H., James D. Wright, and Andy B. Anderson. 1983. *Handbook of Survey Research*. Orlando: Academic Press.

Sherman, S. J. 1980. "On the Self-Erasing Nature of Errors of Prediction." *Journal of Personality and Social Psychology* 39:211-21.

Siemiatycki, Jack. 1979. "A comparison of Mail, Telephone, and Home Interview Strategies for Household Health Surveys." *American Journal of Mental Health* 69: 238-245.

Smith, Eliot R., Nyla Branscombe, and Carol Bormann. 1988. "Generality of the Effects of Practice on Social Judgement Tasks." *Journal of Personality and Social Psychology* 54:385-395.

Sobol, Marion G. 1959. "Panel Mortality and Panel Bias." *Journal of the American Statistical Association* 54(285):52-68.

Sudman, Seymour, and Norman M. Bradburn. 1974. *Response Effects in Surveys: A Review and Synthesis*. Chicago: Aldine.

Tarnai, John, and Don Dillman. 1992. "Questionnaire Context as a Source of Response Differences in Mail and Telephone Surveys." In *Context Effects in Social and Psychological Research*, eds. Norbert Schwarz and Seymour Sudman. New York, NY: Springer-Verlag.

Tourangeau, Roger and Tom W. Smith. 1996. "Asking Sensitive Questions: The Impact of Data

Collection Mode , Question Format, and Question Context." *Public Opinion Quarterly* 60:275-304.

van der Zouwen, Johannes, Wil Dijkstra, and Johannes H. Smit. 1991. "Studying Respondent-interviewer Interaction: The Relationship between Interviewing Style, Interviewer Behavior, and Response Behavior." In *Measurement Errors in Surveys*, ed. Paul Biemer, Robert M. Groves, Lars E. Lyberg, Nancy Mathiowetz, and Seymour Sudman. New York: John Wiley and Sons.

Willson, Victor L. and Richard R. Putnam. 1982. "A Meta-Analysis of Pretest Sensitization Effects in Experimental Design." *American Educational Research Journal* 19(2):249-58.

Wiseman, Frederick. 1972. "Methodological Bias in Public Opinion Surveys." *Public Opinion Quarterly* 36:105-108.

Wright, Debra L., William S. Aquilino, and Andrew J. Supple. 1998. "A Comparison of Computer-assisted Paper-and-pencil Self-administered Questionnaires in a Survey on Smoking, Alcohol, and Drug Use." *Public Opinion Quarterly* 62:331-353.

Zagorsky, Jay and Pat Rhoton. 1999. *Attrition and the National Longitudinal Survey's Women Cohorts*. Center for Human Resource Research, Ohio State University.

Figure 1: Model of Criterion Validity

Figure 2: Structural Equation Model Used to Estimate Item Reliability

Table 1: National Survey Samples, Field Periods, and Response Rates

	OSU Center for Survey Research	Knowledge Networks	Harris Interactive
Pre-election Survey			
Eligible Households	3,500	7,054	12,523
Participating Respondents	1,506	4,933	2,306
Response Rate	43%	25%[a]	NA
Cooperation Rate	51%	31%	18%
Start Date	June 1, 2000	June 1, 2000	July 21, 2000
Stop Date	July 19, 2000	July 28, 2000	July 31, 2000
Post-election Survey			
Eligible Households	1,506	4,143[b]	2,306
Participating Respondents	1,206	3,416	1,028
Response Rate	80%	82%	45%
Start Date	Nov 9, 2000	Nov 8, 2000	Nov 9, 2000
Stop Date	Dec 12, 2000	Nov 21, 2000	Nov 26, 2000

[a]This figure is the product of 89% (the rate at which eligible RDD-sampled telephone numbers were contacted for initial telephone interviews) and 56% (the rate at which contacted households agreed to participate in the initial telephone interview and agreed to join the KN panel) and 72% (the rate at which households that agreed to join the KN panel had the WebTV device installed in their homes) and 70% (the rate at which invited KN panel respondents participated in the survey).

[b]Of the 4,933 who completed all of the first three instruments, 790 members were excluded from assignment to the follow-up survey for the following reasons: (a) temporarily inactive status (being on vacation, health problems etc.), (b) some individuals had been withdrawn from the panel, and (c) some individuals had already been assigned to other surveys for the week of the election.

Table 2: Demographic Composition of Pre-election Samples compared to CPS data

	OSU Center for Survey Research		Knowledge Networks		Harris Interactive		2000 CPS March Supplement
	Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted	
Education							
Some high school	7.0%	17.1%	6.7%	12.3%	2.0%	7.9%	16.9%
High school grad	31.3%	32.7%	24.4%	33.5%	11.8%	36.5%	32.8%
Some college	19.6%	19.8%	32.3%	28.5%	36.6%	26.9%	19.8%
College grad	30.1%	21.7%	26.0%	18.2%	25.8%	19.8%	23.0%
Postgrad work	12.0%	8.6%	10.6%	7.4%	23.7%	9.0%	7.5%
TOTAL	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
N	1504	1504	4925	4925	2306	2250	
Average Error	4.6%	0.5%	7.4%	3.8%	13.9%	4.9%	
Income							
<$25,000	19.0%	19.0%	14.3%	18.0%	12.6%	24.8%	30.5%
$25-50,000	36.9%	37.1%	32.5%	35.3%	32.3%	29.8%	28.3%
$50-75,000	22.0%	22.4%	27.5%	25.8%	25.9%	20.6%	18.2%
$75-100,000	12.9%	13.4%	13.8%	11.9%	14.8%	11.6%	10.1%
$100,000	9.2%	8.1%	11.9%	9.0%	14.5%	13.0%	12.5%
TOTAL	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
N	1138	1138	4335	4335	1976	1917	
Average Error	6.0%	6.4%	6.8%	6.5%	8.6%	2.3%	
Age							
18-24	10.0%	13.5%	7.8%	9.8%	8.0%	14.0%	13.2%
25-34	17.9%	15.3%	19.1%	19.1%	21.2%	18.9%	18.7%
35-44	24.5%	22.7%	25.8%	22.8%	21.5%	21.8%	22.1%
45-54	20.7%	17.8%	23.0%	19.8%	27.9%	20.4%	18.3%
55-64	12.1%	12.4%	12.4%	13.4%	15.5%	10.4%	11.6%
65-74	9.4%	12.5%	7.7%	9.7%	4.8%	12.3%	8.7%
75+	5.5%	5.8%	4.2%	5.5%	1.0%	2.2%	7.4%
TOTAL	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
N	1496	1496	4923	4923	2306	2250	
Average Error	1.7%	1.6%	2.7%	1.5%	4.6%	1.9%	
Race							
White	78.5%	83.3%	86.4%	82.8%	89.6%	81.1%	83.3%
African American	9.7%	11.9%	6.9%	10.0%	3.6%	12.3%	11.9%
Other	11.8%	4.8%	6.7%	7.2%	6.8%	6.6%	4.8%
TOTAL	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
N	1490	1490	4721	4721	2183	2132	
Average Error	4.7%	0.0%	3.3%	1.6%	5.5%	1.5%	
Gender							
Male	45.1%	46.9%	49.2%	49.2%	60.1%	48.2%	48.0%
Female	54.9%	53.1%	50.8%	50.8%	39.9%	51.8%	52.0%
TOTAL	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
N	1506	1506	4910	4910	2306	2250	
Average Error	2.9%	1.1%	1.2%	1.2%	12.1%	0.2%	
AVERAGE ERROR	4.0%	1.9%	4.3%	2.9%	8.7%	2.2%	

Table 3: Demographic Composition of Post-election Samples compared to CPS data

	OSU Center for Survey Research		Knowledge Networks		Harris Interactive		2000 CPS March Supplement
	Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted	
Education							
Some high school	6.6%	17.1%	7.0%	13.5%	1.1%	7.5%	16.9%
High school grad	29.1%	31.6%	25.9%	32.9%	10.9%	39.5%	32.8%
Some college	20.1%	21.1%	31.9%	28.2%	35.5%	27.1%	19.8%
College grad	31.6%	21.7%	24.9%	18.3%	26.8%	17.3%	23.0%
Postgrad work	12.6%	8.5%	10.3%	7.1%	25.8%	8.6%	7.5%
TOTAL	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
N	1201	1201	3404	3404	1040	1040	
Average Error	5.6%	1.0%	6.7%	3.4%	15.1%	6.0%	
Income							
<$25,000	17.1%	17.5%	15.0%	19.9%	10.0%	18.9%	30.5%
$25-50,000	36.9%	37.7%	33.4%	36.3%	32.1%	31.9%	28.3%
$50-75,000	22.4%	22.3%	27.6%	25.4%	27.1%	20.9%	18.2%
$75-100,000	14.4%	14.7%	13.1%	10.9%	15.9%	12.8%	10.1%
>$100,000	9.3%	7.8%	10.8%	7.5%	15.0%	15.5%	12.5%
TOTAL	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
N	917	917	3006	3006	882	882	
Average Error	6.7%	7.2%	6.9%	6.3%	8.3%	4.7%	
Age							
18-24	8.1%	12.9%	5.9%	9.5%	6.3%	15.7%	13.2%
25-34	17.2%	15.9%	18.2%	20.6%	18.7%	17.5%	18.7%
35-44	24.6%	22.4%	24.3%	22.7%	19.6%	22.0%	22.1%
45-54	22.1%	18.2%	22.9%	19.1%	30.5%	19.3%	18.3%
55-64	12.1%	11.7%	14.0%	13.1%	17.6%	11.1%	11.6%
65-74	10.1%	13.1%	9.5%	9.2%	6.4%	12.7%	8.7%
75+	5.7%	5.8%	5.4%	5.7%	0.9%	1.6%	7.4%
TOTAL	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
N	1197	1197	3408	3408	1040	1040	
Average Error	2.4%	1.4%	2.8%	1.6%	5.2%	2.2%	
Race							
White	79.7%	83.2%	87.5%	81.9%	91.2%	81.4%	83.3%
African American	9.0%	11.9%	6.6%	10.3%	2.9%	12.7%	11.9%
Other	11.3%	4.8%	5.1%	7.9%	5.8%	5.8%	4.8%
TOTAL	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
N	1192	1192	4721	4721	1040	1040	
Average Error	4.3%	0.0%	3.3%	2.1%	6.0%	1.2%	
Gender							
Male	44.6%	47.1%	49.8%	48.0%	59.8%	48.8%	48.0%
Female	55.4%	52.9%	50.2%	52.0%	40.2%	51.2%	52.0%
TOTAL	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
N	1203	1203	4910	4910	1040	1040	
Average Error	3.4%	0.9%	1.8%	0.0%	11.8%	0.8%	
AVERAGE ERROR	4.5%	2.1%	4.3%	2.7%	9.3%	3.0%	

Table 4: Indicators of Interest in Politics

	OSU Center for Survey Research		Knowledge Networks		Harris Interactive	
	Unweighted Sample	Weighted Sample	Unweighted Sample	Weighted Sample	Unweighted Sample	Weighted Sample
Political Knowledge Quiz						
Average Percentage of Correct Responses per Respondent	53%	50%	58%	62%	77%	70%
N	1506	1506	4940	4935	2306	2250
Mid-point Selection						
Average Percentage of Midpoint Selection per Respondent	43.2%	43.8%	39.4%	39.5%	34.0%	33.9%
N	1506	1506	4940	4935	2306	2250
Party Identification						
Percentage of Independents	21.8%	23.3%	22.0 %	23.6%	13.1%	13.6%
N	1461	1458	4792	4803	2306	2250
Pre-election Reports of Electoral Participation						
Will Vote in Presidential Election? Yes	86.2%	84.6%	81.5%	78.5%	94.8%	90.7%
No	13.8%	15.4%	18.5%	21.5%	5.2%	9.3%
TOTAL	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
N	1456	1452	4914	4915	2313	2250
Post-election Reports of Electoral Participation						
Usually Voted in Past Elections? Yes	78.7%	74.4%	76.5%	70.2%	90.8%	83.7%
No	17.9%	21.0%	18.5%	22.4%	6.5%	13.3%
Ineligible	3.2%	4.6%	5.0%	7.4%	2.7%	3.0%
TOTAL	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
N	1206	1204	3408	3408	1040	1028
Voted in 2000 Presidential Election? Yes	78.9%	76.5%	77.7%	72.2%	93.8%	90.9%
No	21.1%	23.5%	22.3%	27.8%	6.3%	9.1%
TOTAL	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
N	1206	1205	3408	3406	1040	1028

Table 5: Change in Probability that Respondent will Vote for Gore instead of Bush (Pre-election Vote Choice) if Change from Minimum to Maximum Scale Point in the Predictor Variable

	Unweighted Samples			Weighted Samples		
	CSR	KN	HI	CSR	KN	HI
Clinton Approval: Job	.73**	.85**	.88**	.71**	.84**	.88**
Clinton Approval: Economy	.67**	.78**	.81**	.65**	.78**	.80**
Clinton Approval: Foreign Relations	.65**	.81**	.85**	.62**	.81**	.82**
Clinton Approval: Crime	.54**	.79**	.87**	.56**	.79**	.85**
Clinton Approval: Race Relations	.61**	.80**	.84**	.58**	.81**	.86**
Clinton Approval: Pollution	.46**	.78**	.85**	.47**	.78**	.86**
Past Conditions: Economy	.50**	.67**	.73**	.48**	.68**	.71**
Past Conditions: Foreign Relations	.76**	.86**	.91**	.74**	.86**	.91**
Past Conditions: Crime	.44**	.74**	.79**	.41**	.76**	.71**
Past Conditions: Race Relations	.45**	.84**	.87**	.42**	.83**	.81**
Past Conditions: Pollution	.21**	.65**	.73**	.21**	.65**	.73**
Expectations: Economy	.52**	.53**	.51**	.54**	.52**	.48**
Expectations: Foreign Relations	.44**	.45**	.40**	.48**	.44**	.36**
Expectations: Crime	.47**	.46**	.43**	.50**	.45**	.41**
Expectations: Race Relations	.61**	.62**	.70**	.63**	.59**	.61**
Expectations: Pollution	.60**	.67**	.78**	.60**	.64**	.68**
Candidates' Traits: Moral	.52**	.59**	.59**	.54**	.56**	.53**
Candidates' Traits: Really Cares	.54**	.60**	.68**	.57**	.59**	.63**
Candidates' Traits: Intelligent	.52**	.59**	.67**	.55**	.55**	.57**
Candidates' Traits: Strong Leader	.30**	.32**	.29**	.37**	.31**	.29**
Evoked Emotions: Angry	.56**	.56**	.59**	.58**	.53**	.52**
Evoked Emotions: Hopeful	.41**	.51**	.54**	.45**	.50**	.50**
Evoked Emotions: Afraid	.56**	.55**	.61**	.59**	.52**	.55**
Evoked Emotions: Proud	.44**	.49**	.49**	.48**	.48**	.44**
Party Identification	.94**	.91**	.95**	.91**	.91**	.91**
Political Ideology	.76**	.93**	.94**	.74**	.91**	.90**
Military Spending	.58**	.55**	.72**	.55**	.50**	.68**
Welfare Spending	.53**	.61**	.72**	.42**	.60**	.61**
Help for Black Americans	.61**	.60**	.74**	.56**	.61**	.73**
Gun Control	.58**	.52**	.63**	.55**	.52**	.64**
Pollution by Businesses	.34**	.37**	.53**	.26**	.33**	.54**
Effort to Control Crime	.16*	.17**	.12*	.18**	.23**	.25**
Immigration Restriction	.20**	.19**	.32**	.15*	.17**	.26**
Make Abortion Illegal	.31**	.34**	.48**	.31**	.29**	.40**
Make Abortion Legal	.40**	.36**	.51**	.42**	.32**	.45**
Help Poor Countries Provide For People	.26**	.35**	.39**	.17**	.36**	.37**
Prevent People In Other Countries From Killing Each Other	.27**	.29**	.45**	.24**	.19**	.40**
Prevent Other Governments From Hurting Their Own Citizens	.26**	.27**	.41**	.21**	.26**	.37**
Resolve Disputes Between Other Countries	.20**	.25**	.38**	.17*	.24**	.27**
Prevent Other Countries From Polluting the Environment	.21**	.37**	.50**	.18*	.36**	.43**
Build Missile Defense System	.31**	.38**	.52**	.29**	.32**	.42**
Average Change in Probability	.47	.56	.63	.46	.55	.59

* p<.05; ** p<.01

Table 6: Change in Probability that Respondent will Vote for Gore instead of Bush (Post-election Vote Choice) if Change from Minimum to Maximum Scale Point in the Predictor Variable

	Unweighted Samples			Weighted Samples		
	CSR	KN	HI	CSR	KN	HI
Clinton Approval: Job	.77**	.87**	.93**	.77**	.88**	.86**
Clinton Approval: Economy	.69**	.80**	.86**	.68**	.82**	.79**
Clinton Approval: Foreign Relations	.67**	.83**	.91**	.65**	.83**	.84**
Clinton Approval: Crime	.58**	.80**	.92**	.62**	.78**	.85**
Clinton Approval: Race Relations	.61**	.81**	.88**	.59**	.78**	.85**
Clinton Approval: Pollution	.44**	.78**	.89**	.42**	.76**	.81**
Past Conditions: Economy	.50**	.70**	.74**	.47**	.72**	.70**
Past Conditions: Foreign Relations	.76**	.89**	.94**	.78**	.88**	.94**
Past Conditions: Crime	.49**	.73**	.81**	.48**	.73**	.73**
Past Conditions: Race Relations	.44**	.84**	.93**	.48**	.82**	.94**
Past Conditions: Pollution	.19**	.66**	.78**	.22**	.64**	.80**
Expectations: Economy	.47**	.45**	.46**	.45**	.44**	.46**
Expectations: Foreign Relations	.40**	.39**	.35**	.40**	.37**	.35**
Expectations: Crime	.41**	.40**	.37**	.42**	.37**	.39**
Expectations: Race Relations	.54**	.53**	.62**	.53**	.50**	.57**
Expectations: Pollution	.54**	.56**	.74**	.49**	.53**	.67**
Candidates' Traits: Moral	.45**	.49**	.51**	.44**	.45**	.48**
Candidates' Traits: Really Cares	.47**	.49**	.57**	.47**	.46**	.57**
Candidates' Traits: Intelligent	.45**	.50**	.59**	.43**	.45**	.54**
Candidates' Traits: Strong Leader	.28**	.31**	.25**	.30**	.28**	.27**
Evoked Emotions: Angry	.49**	.47**	.49**	.48**	.44**	.54**
Evoked Emotions: Hopeful	.37**	.40**	.41**	.37**	.38**	.44**
Evoked Emotions: Afraid	.50**	.45**	.51**	.49**	.42**	.52**
Evoked Emotions: Proud	.39**	.41**	.39**	.38**	.39**	.38**
Party Identification	.90**	.91**	.96**	.88**	.90**	.94**
Political Ideology	.81**	.94**	.96**	.79**	.91**	.96**
Military Spending	.62**	.52**	.77**	.60**	.46**	.61**
Welfare Spending	.59**	.61**	.76**	.51**	.61**	.49**
Help for Black Americans	.61**	.61**	.81**	.66**	.63**	.72**
Gun Control	.61**	.59**	.71**	.59**	.61**	.61**
Pollution by Businesses	.33**	.41**	.60**	.27*	.33**	.55**
Effort to Control Crime	.13*	.23**	.20**	.11	.3**	.32**
Immigration Restriction	.21**	.19**	.33**	.15	.19**	-.01
Make Abortion Illegal	.39**	.37**	.56**	.39**	.32**	.48**
Make Abortion Legal	.41**	.36**	.61**	.40**	.33**	.55**
Help Poor Countries Provide For People	.25**	.37**	.43**	.24**	.37**	.14**
Prevent People In Other Countries From Killing Each Other	.31**	.32**	.51**	.31**	.31**	.36**
Prevent Other Governments From Hurting Their Own Citizens	.25**	.30**	.45**	.19**	.29**	.25*
Resolve Disputes Between Other Countries	.15*	.28**	.41**	.13*	.25**	.29**
Prevent Other Countries From Polluting the Environment	.22**	.42**	.57**	.17*	.43**	.51**
Build Missile Defense System	.35**	.36**	.55**	.38**	.32**	.30**
Average Change in Probability	.46	.54	.64	.45	.53	.57

* p<.05; ** p<.01

Table 7: Structural Equation Model Parameter Estimates for Assessing Reliability

Parameter	Indicator	Unweighted			Weighted		
		CSR	KN	HI	CSR	KN	HI
Factor Loadings	$Bush_1$	.61	.77	.85	.58	.74	.82
	$Gore_1$	-.73	-.78	-.86	-.69	-.74	-.80
	$Bush_2$	.72	.78	.88	.68	.78	.91
	$Gore_2$	-.79	-.80	-.86	-.76	-.79	-.83
Error Variances	$Bush_1$	.62	.41	.29	.66	.46	.34
	$Gore_1$	.47	.40	.26	.53	.46	.36
	$Bush_2$	.48	.39	.23	.54	.39	.18
	$Gore_2$	.38	.37	.27	.43	.38	.32

Table 8 Unstandardized Regression Coefficients of Variables Predicting Difference Between Gore and Bush Thermometers.

	Intercom	Computer	z-test
Clinton Approval: Job	.52** (.08)	.88** (.09)	2.95**
	N=166	N=166	
Clinton Approval: Economy	.35** (.12)	.78** (.13)	2.35**
	N=166	N=166	
Clinton Approval: Foreign Relations	.27* (.11)	.79** (.12)	3.2**
	N=166	N=166	
Clinton Approval: Crime	.07 (.10)	.90** (.13)	5.06**
	N=166	N=164	
Clinton Approval: Education	.08 (.10)	.74** (.12)	4.17**
	N=166	N=166	
Clinton Approval: Race Relations	.22 (.12)	.84** (.13)	3.51**
	N=166	N=164	
Clinton Approval: Pollution	-.13 (.10)	.69** (.14)	4.67**
	N=166	N=165	
Clinton Approval: Health Care	.16 (.10)	.83** (.11)	4.32**
	N=166	N=166	
Past Conditions: Economy	.34** (.12)	.40* (.16)	.33
	N=166	N=164	
Past Conditions: Foreign Relations	.29* (.11)	.56** (.15)	1.46
	N=166	N=163	
Past Conditions: Crime	.07 (.11)	.54* (.14)	2.69**
	N=166	N=164	
Past Conditions: Education	.15 (.13)	.48** (.14)	1.75*
	N=166	N=164	
Past Conditions: Race Relations	.04 (.14)	.33* (.17)	1.35
	N=166	N=164	
Past Conditions: Pollution	-.10 (.12)	.55** (.14)	3.65**
	N=166	N=164	
Past Conditions: Health Care	.26* (.10)	.76** (.14)	2.92**
	N=166	N=162	
Expectations: Economy	.56** (.05)	.82** (.05)	3.63**
	N=166	N=164	
Expectations: Foreign Relations	.47** (.05)	.76** (.05)	3.99**
	N=166	N=166	
Expectations: Crime	.41** (.07)	.81** (.06)	4.34**
	N=166	N=166	
Expectations: Education	.52** (.06)	.79** (.05)	3.46**
	N=166	N=166	
Expectations: Race Relations	.54** (.09)	.82** (.07)	2.60**
	N=166	N=166	
Expectations: Pollution	.22** (.08)	.57** (.07)	3.31**
	N=166	N=166	
Expectations: Health Care	.45** (.06)	.76** (.06)	3.66**
	N=166	N=166	
Candidates' Traits: Moral	.50** (.07)	.79** (.07)	2.98**
	N=166	N=166	
Candidates' Traits: Really Cares	.68** (.05)	.84** (.04)	2.37**
	N=166	N=166	
Candidates' Traits: Intelligent	.30** (.08)	.82** (.07)	5.08**
	N=166	N=166	
Candidates' Traits: Strong Leader	.57** (.06)	.76** (.05)	2.48**
	N=166	N=166	
Evoked Emotions: Angry	.74** (.05)	.85** (.05)	1.61
	N=166	N=166	
Evoked Emotions: Hopeful	.64** (.05)	.84** (.04)	3.02**
	N=166	N=166	
Evoked Emotions: Afraid	.70** (.09)	.83** (.07)	1.08
	N=166	N=166	
Evoked Emotions: Proud	.67** (.05)	.88** (.05)	2.99**
	N=166	N=166	
Party Identification	.77** (.09)	1.32** (.10)	4.28**
	N=166	N=166	
Political Ideology	.47** (.10)	.88** (.13)	2.51**
	N=166	N=166	
Military Spending	.30** (.10)	.45** (.14)	.87
	N=166	N=166	
Welfare Spending	.39** (.10)	.61** (.10)	1.51
	N=166	N=166	
Help for Black Americans	.53** (.12)	.65** (.12)	.71
	N=166	N=166	
Gun Control	.23* (.11)	.65** (.15)	2.29*
	N=166	N=166	
Effort to Control Crime	-.12 (.11)	.47* (.19)	2.65**
	N=166	N=165	
Immigration Restriction	.12 (.10)	.30* (.14)	1.02
	N=166	N=166	

* p<.05; ** p<.01; Standard error in parentheses

Figure 1: Model of Criterion Validity

$$\varepsilon_1$$

$$\varepsilon_2$$

| Report of Predictor$_1$ | $\xleftrightarrow{}$ | Report of Vote Choi... |

r_{12}

λ_1

Predictor's Latent Construct $\xleftrightarrow{}$ True Vote Choice
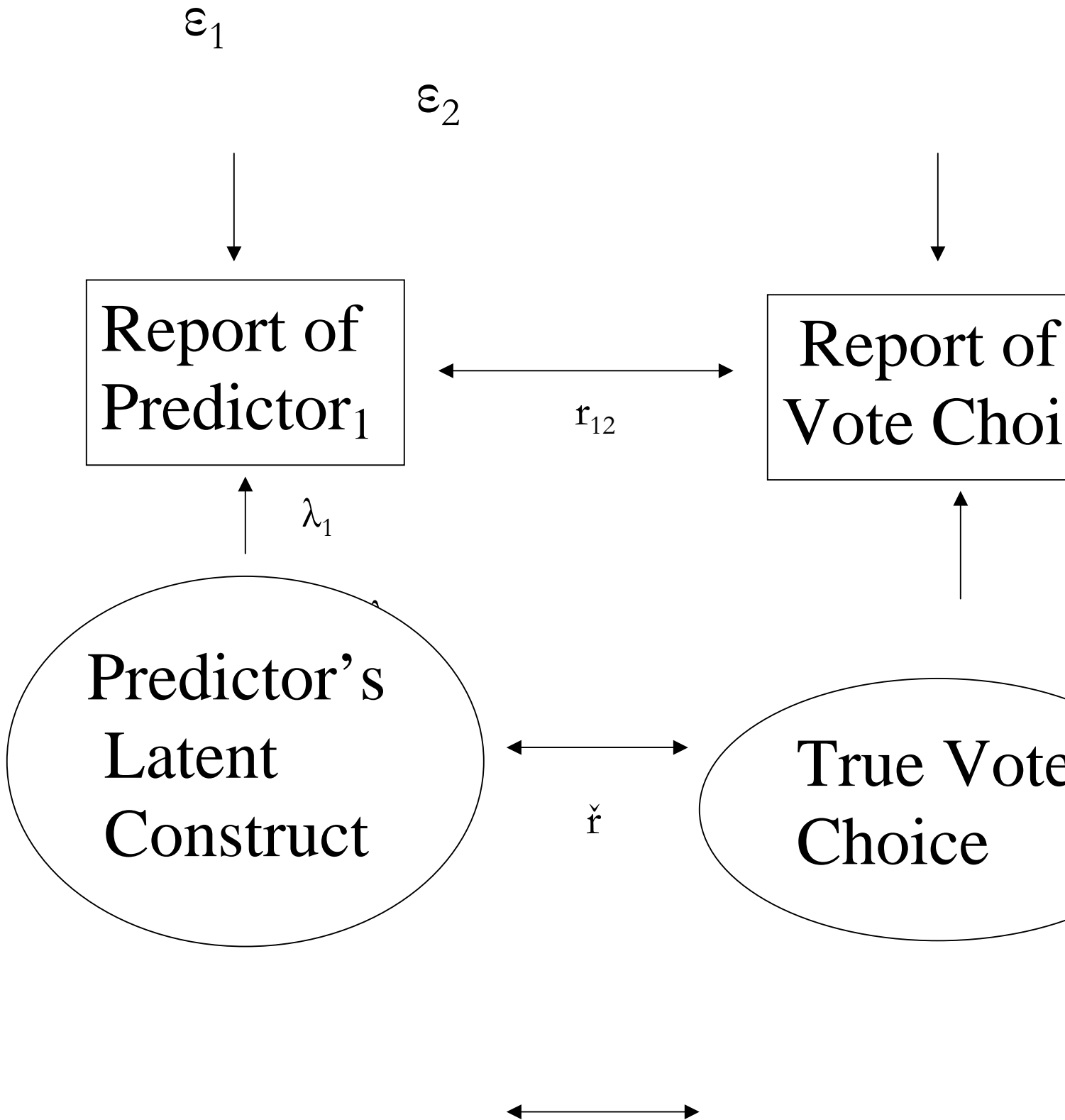
$\check{r}$

$\xleftrightarrow{}$

Figure 2: Structural Equation Model Used to Estimate Item Reliability