Proceedings of Statistics Canada Symposium 2001
Achieving Data Quality in a Statistical Agency: A Methodological Perspective

ARE REPRESENTATIVE INTERNET SURVEYS POSSIBLE?

Tom W. Smith[1]

ABSTRACT

Many types of Web surveys are not based on scientific sampling and do not represent any well-defined population. Even when Web surveys are based on a generalizable sample, it is not known whether they yield reliable or valid results. One way of testing the adequacy of Web surveys is to do experiments that compare Web surveys to well-established, traditional survey modes. This was done when the 2000 General Social Survey of the National Opinion Research Center was compared to a Web survey by Knowledge Networks.

1. INTRODUCTION

Internet surveys come in many varieties. Based on a typology by Couper (2000) the main versions can be characterized as follows:

A. Non-probability
    1. Unrestricted, Self-selection
    2. Restricted, Self-selection
    3. Recruited, Opt-in Panels
B. Probability
    1. Internet Only
        a. Intercepts
        b. List-based
        c. Pre-recruited Panels of Internet Users
    2. General
        a. Mix-mode
        b. Pre-recruited Panels of General Population

Being based on non-probability methods raises a serious barrier to any Internet survey's claims of representativeness and generalizability, although certain practitioners of Recruited, Opt-in Panels such as Harris Interactive argue that their Internet surveys do surmount this obstacle.[2] However, for the purpose of this paper non-probability Internet surveys are not considered as producing representative data.

The probability-based, Internet-only surveys have a solid theoretical basis and have produced some creditable results. But response rates have generally been lower than for other modes and they are naturally restricted to either active Internet users (as in the Intercept surveys) or populations with complete or near complete coverage by email and, usually, Web access (typically college students and certain groups of employees). In the case of Intercept surveys the sample is restricted to those accessing some cooperating Web site. For the List-based

[1]Tom W. Smith, NORC/University of Chicago, 1155 East 60th Street, Chicago, IL, USA, 60615

[2]On these types of Internet surveys in general see Couper, 2000. For examples see Bailey, Foote, and Throckmorton, 1995 and Goeritz and Schumacher, 2000. On the debate over Harris Interactive see Mitofsky, 1999a; 199b; Taylor and Terhanian, 1999a; 1999b; and Taylor, et al., 2001a; 2001b; 2001c. In particular, see the experimental comparisons in Krosnick and Chang, 2001.

surveys the sample depends on access to the master list of email addresses for the target population. The Pre-recruited Panel surveys use some non-Internet methods such as an RDD survey to build a panel of email addresses. Given the decentralized nature of the Internet, existing privacy norms, and other technical barriers, there does not appear to be any way to draw a general sample of all Internet users akin to RDD sampling of telephone households.[3]

For the general population the mix-mode approach can involve people or organizations sampled and contacted by some non-Internet approach for whom a Web survey is one of the offered means to respond to the request for data or an Internet/email based contact from a list using mail, fax, and/or some other mode in addition to the Web survey medium. These may involve surveys that are heavily Internet or those that mostly use other modes with a small Internet component.[4]

The last version of Internet surveys, Pre-recruited Panels of the General Population, draws a sample of the general population using such approaches as RDD and then both recruits respondents into a research panel and supplies them with the Internet conductivity that the panelists need to participate in Internet surveys. It differs from the Internet only, Pre-recruited Panels in that it is not restricted to current Internet users, but instead turns all respondents into Internet users equipped with the same systems.[5]

2. THE EXPERIMENT

An experiment was designed to compare the most promising of the Internet survey procedures, Pre-recruited Panels of the General Population with a high, quality non-Internet survey. This was done by taking a series of questions that appeared on the General Social Survey and placing them on a survey of Knowledge Networks. The 2000 General Social Survey (GSS) is an in-person, multi-stage, area probability sample of adults living in households in the United States conducted by the National Opinion Research Center (NORC), University of Chicago. The 2000 GSS had a sample size of 2817 and a response rate of 70%. The field dates were from early February to the end of May with the highest concentration of completed cases in March. For more details see Davis, Smith, and Marsden, 2001. The Knowledge Networks survey (KNS) was designed and commissioned by NORC and carried out by Knowledge Networks (KN). KN respondents consist of people contacted via RDD surveys (originally conducted by NORC and later by RTI) and recruited into a panel. Recruits are provided with WebTV and access to the Internet via WebTV. The panelists agree to periodically answer surveys sent to them via WebTV. KNS had a sample size of 1413 and a response rate of about 40%.[6] Data collection was March 3-13, 2000. General information on KN can be found at www. knowledgenetworks.com.

The items chosen for comparison were the items that appeared first on the GSS. This controlled for possible

[3]For a general discussion see Couper, 2000. For specific examples see Anderson and Gansneder, 1995; Bates, 2001; Burr, Levin, and Beecher, 2001; Couper, Blair, and Triplett, 1999; Dommeyer and Moriarty, 2000; Jones and Pitt, 1999; Liu, Rosen, and Stewart, 2001; Manfreda, Vehovar, and Batagelj, 2001; Pew, 1999; Ramirez, 2000; Schaeffer and Dillman, 1998; Schuldt and Totten, 1994; Sheehan, 2001; Stone, Vespia, and Kanz, 2000; and Tse, et al., 1995.

[4]See Couper, 2000; Griffin and Holbert, 2001; Liu, Rosen, and Stewart, 2001; Manfreda, Vehovar, and Batagelj, 2001; Ramirez, Sharp, and Foster, 2000; and Zhang, 1999.

[5]On these Internet surveys see Couper, 2000; Greenberg and Rivers, 2001; Kenyon, Couper, and Tourangeau, 2001; and Krosnick and Chang, 2001.

[6]Knowledge Networks reported that the recruitment rate into the panel was 57% and that the response of the panelists selected for the KNS was 71%, giving an overall response rate of 40%. However, some other general information on Knowledge Networks methodology indicates that the true, overall response rate might be in the mid-30s.

context effects from prior items. The first set of items were 16 questions on spending priorities (full question wordings available from author). Respondents received one of two versions of these items. The standard/verbose version or the variant/terse version. This question wording split has been carried out on the GSS for many years (Smith, 1987; Rasinski, 1989). This was followed by a) an item on expectations of war, b) an item on school prayer, and c) five items on gender roles.

3. RESULTS

Table 1 (available from author) shows that there is very little difference between the weighted and unweighted distributions on KNS. A similar pattern exists for the GSS. The subsequent analysis uses the weighted KNS and GSS data, but results would have been very similar with unweighted data.

Table 2 (available from author) compares the KNS and GSS distributions with Don't Knows (DKs) included. KNS uniformly registered more DKs than the GSS did. For the 16 standard-wording spending items the average DK levels were 14.4% on the KNS and 6.1% on the GSS and for the variant-wording spending items the average DK levels were 13.9% for KNS and 6.1% for the GSS. In both cases the DK ratios were 2.3:1. For the seven non-spending items the average levels were 8.6% for KNS and 3.8% for the GSS with a DK ratio again of 2.3:1.

The direction of the DK differences are easily explained by a difference in the format between KNS and the GSS. On KNS DKs were explicitly coded responses that appeared with the other response options. On the GSS DKs were a precoded, but unread, response. By making DKs an explicit response presented on an equal basis with other responses, KNS facilitated the selection of DKs. Alternatively, KNS could have excluded DKs as a response option. This would have produced an opposite DK effect, KNS would have had fewer DKs than the GSS. A third option which would probably produce a closer match to the precoded-but-unread format would have been to tell respondents at the beginning of the survey and/or at the bottom of each screen that some special key could be hit to record a DK response. This approach was not part of Knowledge Networks standard repertoire.

While the direction of the DK effect was consistent and the effect was quite stable on average (2.3:1), there was considerable variation in the DK ratios for individual items. For the spending items the ratio ranged from 1.7:1 to 5.0:1. For the seven non-spending items the ratios ranged from 1.3:1 to 3.1:1. On the spending items it is unclear why the ratios are high for some items and low for others, but at least some of the variation of the non-spending items appears explicable. The lowest ratio is on the one gender-role item using a five-point agree/disagree scale with an explicit mid-point (neither agree nor disagree). Compared to four gender-role items that used a four-point agree/disagree scale with no middle option, the five-point scale reduced the level of DKs for both KNS and the GSS and reduced the DK ratio. The average DK levels for the four-point items were 4.6% for KNS and 2.3% for the GSS with a ratio of 2.0:1 vs. levels for the five-point item of respectively 2.7 and 2.0 and a ratio of 1.35:1. However, while this five-point format minimizes DK differences, it produces a very large difference on the added middle category itself, 30.5% on KNS and 17.5% on the GSS.

On the 32 spending items with the DKs excluded, many distributions are usually similar (e.g. with 24 comparisons showing differences of five percentage points or less) and the overall rankings of spending priorities are highly similar. But a number of items (cities, drugs, Blacks, and Welfare) show large differences. For example, the proportion saying that there was too little spending on the standard drug wording varied by 18.1 percentage points and the variant drug wording differed by 8.5 percentage points. Similarly, support for more spending for Blacks differed by 13.4 percentage points on the standard item and 11.9 percentage points on the variant wording.

In addition, there were also consistent differences in direction. On the standard items in 13 of 16 comparisons

KNS respondents were more likely to say that too much was being spent. On the variant items this was true of 15 of 16 comparisons. While most of the differences were not statistically significant, for a quarter of the items there were reliable differences for one or both of the versions. All of the four reverse patterns were very small and statistically insignificant.

Among the eight non-spending items three notable differences occur between KNS and the GSS. First, while differences on the five gender-role items are small when the DKs are excluded and just the agrees and disagrees are compared, there is a consistent difference in the selection of strongly agree or strongly disagree vs. just agree or disagree. In all instances KNS gets more mentions in the strongly category and fewer mentions in the unmodified category than the GSS does. The pattern shows up for both agree and disagree, but is more pronounced in the former case. For example, on preschoolers suffering if the mother works 14.7% strongly agree on KNS and 10.0% of the GSS (-4.7 percentage points), but 29.9% agree on KNS vs. 36.6% of the GSS (+6.7 percentage points).

Second, the largest difference on all non-spending comparisons is on the Supreme Court's decision outlawing prayers in school. This decision is approved of by 55.9% on KNS and 39.0% on the GSS (-16.9 percentage points). This item is known to be confusing to some people. There is a tendency to confuse approving of the Court's banning of school prayers with approving of school prayers themselves. It is likely that KNS and the GSS differ at least in part because of whether this confusion is greater or lesser. This was indirectly assessed by looking at how the items correlated with certain criterion variables. The analysis (available from author) supports the idea that the items may have been interpreted differently across the surveys and that there might be more reversed responses in KNS than in the GSS.

Third, as noted above in the discussion of DK levels, the five-point, men-overworking item shows large differences on the selection of the middle-response option.

4. CONCLUSION

Most existing varieties of Internet surveys either do not yield representative, generalizable results or do so only for very restrictive populations. The one Internet survey version that is based on probability sampling and covers the general population is Pre-Recruited Panels of Internet Users. To examine this most promising of Internet survey approaches an experiment was designed comparing a KN survey and the 2000 GSS. While many comparisons showed similar results from KNS and the GSS, a number of notable differences did occur.

First, KNS systematically produced more DKs than the GSS did. This was probably largely due to differences in format. However, although the DK effect is consistent across groups of items, it varies quite a bit across individual items. While it might be difficult for a Web survey to closely reproduce the DK levels associated with the common, precoded-but-unread approach for handling DKs on non-Internet surveys, there is no reason to believe that either DK level is more valid than the other.

Second, for reasons that are less obvious, KNS produces more extreme responses to agree/disagree scales than the GSS does. Since the GSS did not use a showcard, this could be a difference between a visual and an oral medium (Kenyon, Couper, and Tourangeau, 2001). Or the fact that more key strokes are needed to select answers lower on the Internet response scale may explain some of the favoritism for strongly agree vs. agree. While this effect seems to have little effect on distributions once the categories are collapsed, it may have a systematic impact on relationships when the uncollapsed scales are utilized.

Third, while most differences were small, KNS showed less support for increased spending in 28 of 32 comparisons. This may result in part from the higher DKs on KNS, if the extra DKs were mostly showing up as

pro-spending on the GSS.

Finally, a number of notable differences also occur on particular items. Among the spending items large differences appear on cities, drugs, Blacks, and welfare. While it is uncertain why these items showed large discrepancies, they all dealt with the problems of the urban, underclass. It may be that the interpersonal nature of the GSS interview increases support for such spending. Among the non-spending items the one large difference was on school prayer. This item may have been misunderstood more frequently on KNS than the GSS.

While KNS and the GSS agreed in most comparisons, there were a notable number of systematic differences (on DK level, agree/ disagree scales, and spending levels) and several large differences on specific items. This indicates that it cannot automatically be expected that even Internet surveys based on probability samples and general populations will produce results equivalent to those from non-Internet surveys. Internet surveys intrinsically differ from standard, non-Internet surveys in format and respondent-demand characteristics and will often differ on other characteristics such as population coverage and response rate. These factors will usually combine together to produce notable differences between Internet and non-Internet surveys. Until the differences in the error structures of these different forms of surveying are better understood and can be modeled and minimized, results from the different survey modes are likely to be common and notable.

REFERENCES

Anderson, Susan E. and Gansneder, Bruce M. (1995), "Using Electronic Mail Surveys and Computer-Monitored Data for Studying Computer-Mediated Communication Systems," *Social Science Computer Review*, 13, pp. 33-46.

Bailey, Robert D.; Foote, Winona E.; and Throckmorton, Barbara (2000), "Human Sexual Behavior," in by Michael H. Birnbaum et al. (eds.) *Psychological Experiments on the Internet*, San Diego: Academic Press.

Bates, Nancy (2001), "Internet versus Mail as a Data Collection Methodology from a High-Coverage Population," paper presented at the American Association for Public Opinion Research, Montreal, Canada.

Burr, Michele A.; Levin, Kerry Y.; and Beecher, Angela (2001), "Examining Web vs. Paper Mode Effects in a Federal Government Customer Satisfaction Study," paper presented at the American Association for Public Opinion Research, Montreal, Canada.

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

Couper, Mick P.; Blair, Johnny; and Triplett, Timothy (1999), "A Comparison of Mail and E-mail for a Survey of U.S. Statistical Agencies," *Journal of Official Statistics*, 15, pp. 39-56.

Davis, James A.; Smith, Tom W.; and Marsden, Peter V. (2001), *General Social Survey, 1972-2000: Cumulative Codebook*. Chicago: NORC.

Dommeyer, Curt J. and Moriarty, Eleanor (2000), "Comparing Two Forms of an Email Survey: Embedded vs. Attached," *International Journal of Marketing Research*, 42, pp. 39-50.

Goeritz, Anja S. and Schumacher, Joerg (2000), "The WWW as a Research Medium: An Illustrative Survey on Paranormal Belief," *Perceptual and Motor Skills*, 90, pp. 1195-1206.

Greenberg, Anna and Rivers, Douglas (2001), "Pioneer Days: The Promise of Online Polling," *Public Perspective*, 12, pp. 40-41.

Griffin, Elizabeth K. and Holbert, Heather C. (2001), "A Feasibility Evaluation of a Web-Based Demographic Survey," paper presented at the American Association for Public Opinion Research, Montreal, Canada.

Jones, R. and Pitt, N. (1999), "Health Surveys in the Workplace: Comparison of Postal, Email, and World Wide Web Methods," *Occupational Medicine*, 49, pp. 556-559.

Kenyon, Kristin; Couper, Mick; and Tourangeau, Roger (2001), "Picture This! An Analysis of Visual Effects in Web Surveys," paper presented at the American Association for Public Opinion Research, Montreal, Canada.

Krosnick, Jon A. and Chang, LinChiat (2001), "A Comparison of the Random Digit Dialing Telephone Survey Methodology with Internet Survey Methodology as Implemented by Knowledge Networks and Harris Interactive," unpublished report.

Liu, Kaiya; Rosen, Jeff; and Stewart, Eric (2001), "Validity Issues in Web Derived Survey Data," paper presented at the American Association for Public Opinion Research, Montreal, Canada.

Manfreda, Katja Lozar; Vehovar, Vasja; and Batagelj, Zenel (2001), "Web Versus Mail Questionnaire for an Institutional Survey," in A. Westlake et al. (eds.) *The Challenge of the Internet*, Association for Survey Computing.

Mitofsky, Warren J. (1999a), "Miscalls Likely in 2000," *Public Perspective*, 10, pp. 42-43.

Mitofsky, Warren J. (1999b), "Pollsters.Com," *Public Perspective*, 10, pp. 24-26.

Pew Research Center for the People and the Press (1999), "A Survey Methods Comparisons: Online Polling Offers Mixed Results."

Ramirez, Carl; Sharp, Kevin; and Foster, Luis (2000), "Mode Effects in an Internet/Paper Survey of Employees," paper presented to the American Association for Public Opinion Research, Portland.

Rasinski, Kenneth A. (1989), "The Effects of Question Wording on Public Support for Government Spending," *Public Opinion Quarterly*, 53, pp. 388-396.

Schaeffer, David R. and Dillman, Don A. (1998), "Development of a Standard E-Mail Methodology: Results of an Experiment," *Public Opinion Quarterly*, 62, pp. 378-397.

Schuldt, Barbara A. and Totten, Jeff W. (1994), "Electronic Mail Vs. Mail Survey Response Rates," *Marketing Research*, 6, pp. 36-39.

Sheehan, Kim (2001), "E-mail Survey Response Rates: A Review," *JCMC*, 6, pp. 1-19.

Smith, Tom W. (1987), "That Which We Call Welfare by Any Other Name Would Smell Sweeter: An

Analysis of the Impact of Question Wording on Response Patterns," *Public Opinion Quarterly*, 51, pp. 75-83.

Stone, Gerald L.; Vespia, Kristin M.; and Kanz, Jason E. (2000), "How Good Is Mental Health Care on College Campuses?" *Journal of Counseling Psychology*, 47, pp. 498-510.

Taylor, Humphrey; Brenner, John; Overmeyer, Cary; Siegel, Jonathan W.; and Terhanian, George (2001a), "The Record of Internet-based Opinion Polls in Predicting the Results of 72 Races in the November 2000 US Elections," *Journal of the Marketing Research Society*, 43, pp. 127-136.

Taylor, Humphrey; Brenner, John; Overmeyer, Cary; Siegel, Jonathan W.; and Terhanian, George (2001b), "Touchdown! Online Polling Scores Big in November 2000," *Public Perspective*, 12, pp. 38-39.

Taylor, Humphrey; Brenner, John; Overmeyer, Cary; Siegel, Jonathan W.; and Terhanian, George (2001c), "Using Internet Polling to Forecast the 2000 Election," *Marketing Research*, 13, pp. 26-30.

Taylor, Humphrey and Terhanian, George (1999a), "Heady Days Are Here Again: Online Polling Is Rapidly Coming of Age," *Public Perspective*, 10, pp. 20-23.

Taylor, Humphrey and Terhanian, George (1999b), "No Witchcraft Here," *Public Perspective*, 10, pp. 42-43.

Tse, Alan C.B. et al. (1995), "Comparing Two Methods of Sending Out Questionnaires: E-mail versus Mail," *Journal of the Market Research Society*, 37, pp. 441-446.

Zhang, Yin (1999), "Using the Internet for Survey Research: A Case Study," *Journal of the American Society for Information Science*, 51, pp. 57-68.