Proceedings of Statistics Canada Symposium 2014
Beyond traditional survey taking: adapting to a changing world

Measurement Properties of Web Surveys

Roger Tourangeau[1]

Abstract

Web surveys have serious shortcomings in terms of their representativeness, but they appear to have some good measurement properties. This talk focuses on the general features of web surveys that affect data quality, especially the fact that they are primarily visual in character. In addition, it examines the effectiveness of web surveys as a form of self-administration. A number of experiments have compared web surveys with other modes of data collection. A meta-analysis of these studies shows that web surveys maintain the advantages of traditional forms of self-administration; in particular, they reduce social desirability bias relative to interviewer administration of the questions. I conclude by discussing some likely future developments in web surveys—their incorporation of avatars as "virtual interviewers" and the increasing use of mobile devices (such as tablet computers and smart phones) to access and complete web surveys.

Key Words: Measurement error; Mode effects; Visual features of surveys.

1. Introduction

Web surveys have several important features that may reduce measurement error relative to other methods of data collection:

1) The primary channel for presenting the questions is visual, allowing the incorporation of photographs, video clips, etc.;

2) The questions are administered by the computer rather than by an interviewer, a feature that may reduce social desirability bias in the answers and that offers additional advantages as well;

3) The survey can be interactive, providing various types of help to the respondent and routing the respondents to the appropriate questions;

4) The respondent controls the pace of the survey questions and can easily reread the questions, features that can reduce the cognitive burden of answering.

In this talk, I focus on the first and third of these features.

2. Visual Heuristics for Web Surveys

A potentially desirable feature of web surveys is their ability to present visual information, including still images or videos. Images are likely to attract attention from respondents and, when they are presented alongside the questions, they are likely to shape respondents' interpretation of the questions. Couper, Tourangeau, and Kenyon (2004) conducted an experiment that varied the content of photographs that accompanied each of six survey items. The items asked respondents how often they did various things, such as attending sporting events or going on overnight trips, during the previous year. The photographs presented with the questions depicted an instance of the category of

[1]Roger Tourangeau, Westat, 1600 Research Boulevard, Rockville, Maryland, 20905, USA, RogerTourangeau@Westat.com

interest. The pictures were chosen to represent either low or high frequency examples of the target category (e.g., a major league baseball game or a Little League game). Respondents who got high frequency instances of the target category tended to report higher behavioral frequencies than those who got the low frequency instances. A later study by Couper and colleagues also demonstrated visual context effects, this time on respondents' judgments of their health (Couper, Conrad, and Tourangeau, 2007).

Spacing and other visual cues may also influence how respondents interpret their task or the intended meaning of response scales. Tourangeau and his colleagues (Tourangeau, Couper, and Conrad, 2004, 2007) propose that respondents follow four interpretive heuristics in assigning meaning to the response scales in web surveys. The four heuristics are:

- Middle means typical or central;
- Left and top mean first;
- Near means related; and
- Like (in appearance) means close (in meaning).

According to the first heuristic, the visual midpoint of a scale plays an important role in establishing the meaning of the scale points. The visual midpoint of a bipolar scale will, according to the heuristic, be taken to represent the conceptual midpoint of the dimension of judgment — that is, the neutral point of the scale. When the underlying dimension is unipolar, respondents may assume that the visual midpoint represents the most typical value (the population median or mode). The second heuristic ("Left and top mean first") refers to the expectation that the response options follow some logical order. When the options are displayed horizontally, respondents expect the leftmost option to represent one extreme and the rightmost option the other and the remaining options to proceed in conceptual order from left to right. When the options are displayed vertically, respondent expect the top and bottom options to represent the two extremes and the remaining options to proceed in order from top to bottom. Tourangeau and his colleagues (Tourangeau et al., 2004, Experiment 4) find respondents took much longer to answer the questions when the order of the options was inconsistent with these expectations than when they were consistent with them.

The "Near means related" heuristic refers to the tendency for respondents to infer a conceptual link between two items based on their physical proximity. In a study examining this heuristic, Tourangeau, Couper, and Conrad (2004, Experiment 6) compared three methods for presenting a series of related attitude items. They presented the items a) in a grid on one screen, b) in two grids on different screens, or c) as individual items on adjacent screens. They predicted (and found) the highest correlations among the items were when all of them were presented in the same grid and the lowest when each item was displayed on a separate screen (see also Couper, Traugott, and Lamias, 2001). When the items were placed in a grid together, respondents seemed to expect them to be similar in meaning and they tended to give them more consistent ratings. The fourth heuristic ("Like in appearance means close in meaning") refers to the tendency for respondents to infer conceptual similarity between two items or two response options based on their similarity in appearance. For example, when the two ends of the response scale are shades of the same hue, respondents may infer that the two extremes are closer conceptually than when the two ends of the scales are shades of different hues. Based on this inference, respondents' answers can shift (Tourangeau et al., 2007).

Figure 2-1 below provides an illustration of the impact of the first heuristic. In the top version of the scale, the conceptual midpoint ("Even chance") coincides with the visual midpoint. In the bottom version, conceptual midpoint is to the right of the visual midpoint. We (Tourangeau, Couper, and Conrad, 2004) conducted an experiment comparing the two versions. With the bottom version of the scale, respondents were significantly more likely to pick one of the low-probability scale points ("Possible," "Unlikely," or "Impossible").

Figure 2-1
Displacing the Conceptual Midpoint of a Scale.

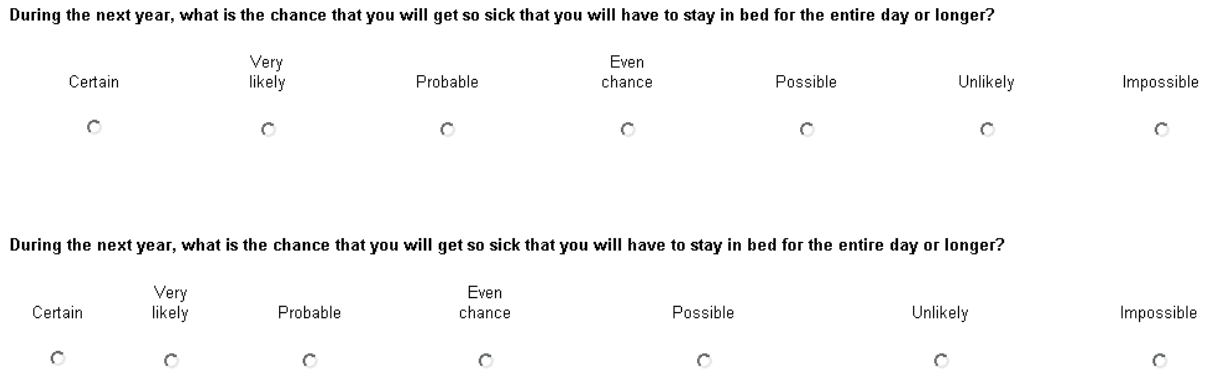
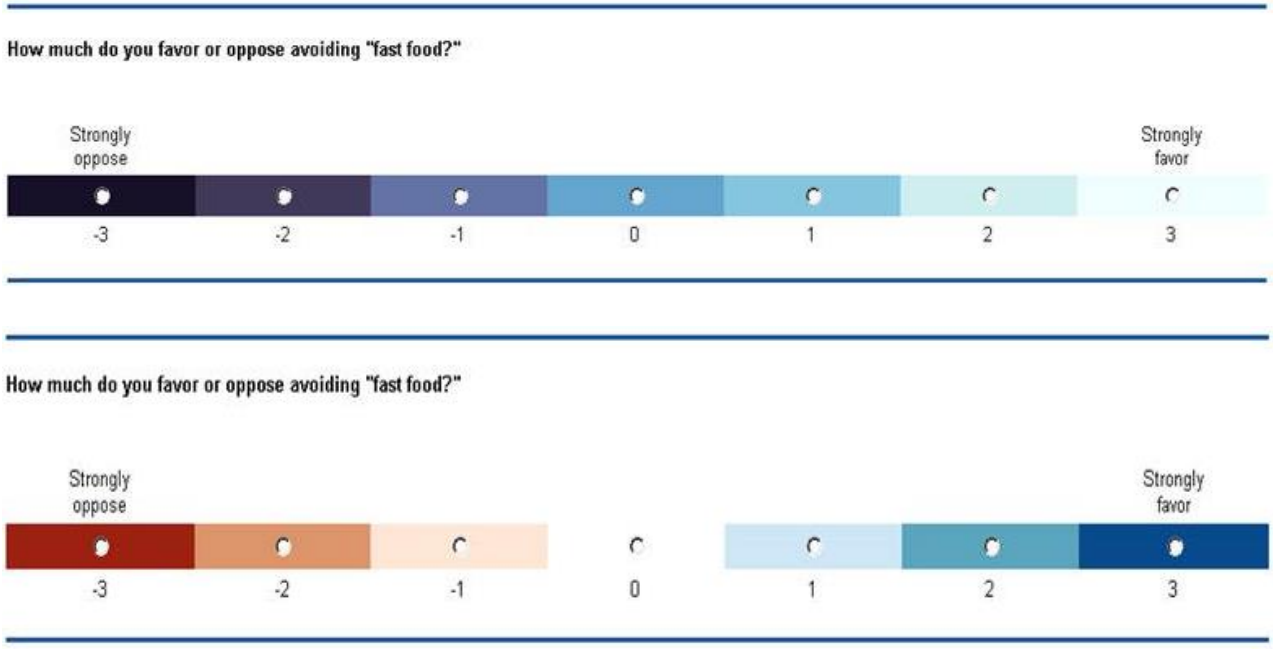


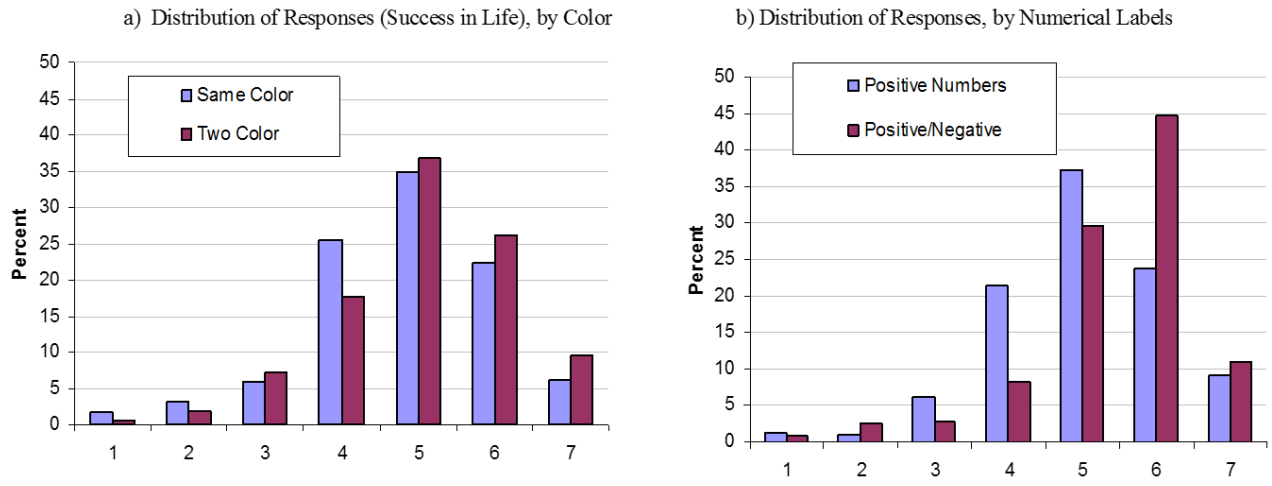
Figure 2-2 below illustrates the "Like (in appearance) means close (in meaning)" heuristic. In two experiments, the respondents received either scales which where shaded with colors of a single hue (as in the top panel of the figure) or with different hues at the two ends of the scale.

Figure 2-2
Single versus Two-color Shading of Scale Points.



The experiment also varied the labeling of the scale points. Some respondents received verbal labels only for scale endpoints, as in the examples in the figure. In addition, within this condition, we varied how the scale points were numbers. For some respondents, the scale points were numbered from 1 to 7; for others, the numbers ran from -3 to +3 (again, as illustrated in the figure); for a third group, the scale points were not numbered. A fourth group got fully labelled scales—that is, every scale point had a verbal label. We predicted that respondents might use the color as cue to the intended meaning of the scale points, just as they use the numbers (as Schwarz et al., 1991, demonstrates). Figure 2-3 shows the results from the experiment. The two-color version of the response scale pushes the answers to the right (panel a), though to a lesser extent than the scale using both positive and negative numbers (panel b).

Figure 2-3
Response Distributions, By Shading of the Scale Points (left graph) and Numerical Labels (right graph).



3. Web Surveys as Form of Self-Administration

One of the most important differences among survey modes is whether or not an interviewer asks the questions and records the answers. A substantial body of evidence shows that respondents are more willing to make potentially embarrassing revelations in surveys when an interviewer is *not* involved and, instead, the respondents fill out paper questionnaires or answer questions that the computer put directly to them (Tourangeau and Yan, 2007). Self-administration — whether via the computer or a paper questionnaire — elicits more accurate answers to sensitive questions about such topics as illicit drug use, alcohol consumption, and voting. Of course, a key feature of online surveys is that no interviewer is involved.

A study by Kreuter, Presser, and Tourangeau (2007) demonstrates this advantage of an Internet survey over other methods of data collection. Kreuter and her colleagues compared reporting errors under three modes of data collection — CATI, IVR, and an online survey. The investigators were able to validate responses to some of the survey items against external records data. For all four of the socially undesirable behaviors that they examined (things like getting an unsatisfactory grade in a class or going on academic probation), the web survey yielded the lowest rate of reporting errors, lower than both CATI (which features interviewer administration) and IVR (which does not). Chang and Krosnick (2009) also looked at the impact of online survey administration on the rate of socially desirable responding. They found that white respondents were more likely to say that the government should provide help to black Americans when the questions were administered by telephone interviewers than when the same questions were administered in a web survey; they interpret these findings to mean that respondents are more willing to give truthful (but socially undesirable) answers in a web survey. Some web surveys feature "avatars" (on-screen representations of the researcher or of an interviewer) or other "humanizing" touches; to date, these features do not appear to heighten social desirability biases (Tourangeau, Couper, and Steiger, 2003). Still, such enhancements to text should probably be avoided when social desirability may be a concern. In addition, "gendered" interfaces (e.g., a web questionnaire that displays a photograph of the female investigator) may influence responses to questions about gender-related topics (e.g., questions about women in the workplace; Tourangeau et al., 2003; see also Fuchs, 2009).

Tourangeau, Conrad, and Couper (2013) carried out a meta-analysis of the results of a number of experiments comparing web surveys with other modes of data collection. In total, the studies included reported 223 mode comparisons, 160 involving web and paper questionnaires. Table 3-1 shows the key results from their meta-analysis; it presents the average effect sizes (typically log odds ratios) for each of the studies.

Table 3-1.
Effects Sizes by Study and Mode Comparison

Comparison	Study	Sample Sizes	Mean Effect Size	Standard Error
Web vs. Mail/Paper	Bälter et al. (2005)	Mail: 188 Web: 295	0.054	0.309
	Bason (2000)	Mail: 204 Web: 115	-0.168	0.129
	Bates and Cox (2008)	Mail: 73 Web: 64	-0.014	0.180
	Eaton et al. (2010)	Mail: 1,729 Web: 3,498	0.070	0.012
	Denscombe (2006)	Mail: 267 Web: 69	-0.256	0.120
	Knapp & Kirk (2003)	Mail: 174 Web: 57	-0.077	0.119
	Link & Mokdad (2005a, 2005b)	Mail: 836; 804-820 Web: 1,143; 948-1,139	0.068	0.039
	McCabe et al. (2002, 2006) McCabe (2004)	Mail: 1,412 Web: 2,194	0.006	0.019
Web vs. Telephone	Bason (2000)	Phone: 161 Web: 115	-0.503	0.132
	Chang & Krosnick (2007)	Phone Pre-Election: 1,456 Web Pre-Election: HI: 2,313 KN: 4,914 Phone Post-Election: 1,206 Web Post-Election: HI: 1,040 KN: 3,408	0.172	0.035
	Knapp & Kirk (2003)	Phone: 121 Web: 57	0.193	0.126
	Kreuter et al. (2009)	Phone: 320 Web: 363	0.157	0.060
	Link & Mokdad (2005a, 2005b)	Phone: 2,072; 2,066-2,070 Web: 1,143; 948-1,139	0.026	0.031
Web vs. IVR	Bason (2000)	IVR: 128 Web: 115	0.108	0.143
	Kreuter et al. (2009)	IVR: 320 Web: 363	0.081	0.060

Note: Adapted from Tourangeau, Conrad, and Couper (2013). HI refers to the Harris Interactive panel, KN to the Knowledge Network panel; IVR refers to interactive voice response.

Tourangeau and his co-authors drew two main conclusions from the results. First, web data collection appears to yield more reports of sensitive information than interviewer-administered telephone surveys do (Chang and Krosnick, 2007; Kreuter, Presser, and Tourangeau, 2008; Link and Mokdad, 2005a, 2005b; but see Bason, 2000, for an apparent exception). Across the six studies comparing the two modes, the overall effect size was 0.088, which was not quite significant (t=1.69, df =7), in part due to the study reported by Bason. If that study is dropped, the mean effect size for telephone-web comparisons rises to .105 (with a standard error of .052). At least one study

(Kreuter et al., 2008) shows that the increase in reporting represents an increase in accuracy These findings are in line with prior analyses of the benefits of self-administration for eliciting potentially embarrassing information (Tourangeau and Yan, 2007).

Second, on-line administration seems to offer only a small advantage in reporting over other forms of self-administration. For example, the overall mean effect size for the web-paper comparisons was 0.030, with a standard error of 0.023.

Apart from reduced social desirability biases and interviewer effects, web surveys offer all the usual advantages of automation, including the ability to tailor questions based on pre-loaded information or answers given to earlier items, automated skips and routing, edit checks that ensure the answers are within some predetermined range, randomization of the order of the questions or of the response options, and so on. These features mean that web surveys can administer highly complex questionnaires.

4. Conclusions and Future Directions

Web surveys are still evolving and some of the conclusions presented here are likely to require modification within a few years. The coverage of the general population will almost certainly continue to increase and researchers will doubtless find ways to improve the response rates to web surveys. Even so, it seems unlikely that web surveys will approach the coverage of telephone or face-to-face surveys any time soon. The coverage problems of web surveys are compounded by the difficulties in selecting a representative sample of web users, which have forced many web surveys to rely on samples of self-selected volunteers instead. Given the seriousness of these problems, it is hardly surprising that web surveys do not line up well with general population figures or that weighting adjustments do not usually remove all the discrepancies.

Still, when the representativeness of the respondents is not an important requirement, web surveys have a lot to offer. They can be much cheaper than other forms of data collection and, compared to other types of non-probability samples, web panels can provide a degree of diversity among the respondents. Indeed, the large web panels allow researchers to identify the members of relatively rare subgroups. And web surveys seem to have good measurement properties; the data from web surveys can be accurate, reliable, and valid, perhaps more so than those obtained though the more traditional modes of data collection.

The effort to make Web surveys more human-like may reduce some of these measurement advantages. In three of four comparisons, Fuchs (2009) found that female respondents were significantly more likely to reveal sensitive sexual information (such as whether they had ever had an STD) to a female virtual interviewer than to the male virtual interviewer; pattern was less clear for the male respondents. It seems likely that the more virtual interviewers resemble live ones, the more likely it is that they bring back social desirability and other interviewer effects.

References

Bälter, K.A., Bälter, O., Fondell, E., and Lagaross, Y. T. (2005), "Web-based and Mailed Questionnaires: A Comparison of Response Rates and Compliance", *Epidemiology*, 16, pp. 577-579.

Bason, J. J. (2000), "Comparison of Telephone, Mail, Web, and IVR Surveys of Drugs and Alcohol Use among University of Georgia Students", paper presented at the 55th Annual Conference of the American Association for Public Opinion Research, Portland, OR.

Bates, S. C., and J. M. Cox (2008), "The Impact of Computer versus Paper-Pencil Survey, and Individual versus Group Administration, on Self-Reports of Sensitive Behaviors", *Computers in Human Behavior*, 24, pp. 903-916.

Chang, L., and J. A. Krosnick (2009), "National Surveys via RDD Telephone Interviewing versus the Internet: Comparing Sample Representativeness and Response Quality", *Public Opinion Quarterly*, 73, pp. 641-678.

Couper, M.P., F. G. Conrad, and R. Tourangeau (2007), "Visual Context Effects in Web Surveys", *Public Opinion Quarterly*, 71, pp. 91-112.

Couper, M. P., R. Tourangeau, and K. Kenyon (2004), "Picture This! Exploring Visual Effects in Web Surveys", *Public Opinion Quarterly*, 68, pp. 255-266.

Denscombe, M. (2006), "Web-Based Questionnaires and the Mode Effect: An Evaluation Based on Completion Rates and Data Contents of Near-Identical Questionnaires Delivered in Different Modes", *Social Science Computer Review*, 24, 246-254.

Eaton, D. K., N. D. Brener, L. Kann, M. M. Denniston, T. McManus, T. M. Kyle, A. M. Roberts, K. H. Flint, and J. G. Ross (2010), "Comparison of Paper-and-Pencil versus Web Administration of the Youth Risk Behavior Survey (YRBS): Risk Behavior Prevalence Estimates", *Evaluation Review*, 34, pp. 137-153.

Fuchs, M. (2009), "Gender-of-Interviewer Effects in a Video-Enhanced Web Survey: Results from a Randomized Field Experiment", *Social Psychology*, 40, pp. 37–42.

Kreuter, F., S. Presser, and R. Tourangeau (2008), "Social Desirability Bias in CATI, IVR and Web surveys: The Effects of Mode and Question Sensitivity", *Public Opinion Quarterly*, 72, pp. 847-865.

Link, M. W., and A. H. Mokdad (2005a), "Effects of Survey Mode on Self-Reports of Adult Alcohol Consumption: A Comparison of Mail, Web, and Telephone Approaches", *Journal of Studies on Alcohol*, 66, pp. 239-245.

Link, M. W., and A. H. Mokdad (2005b), Alternative modes for health surveillance surveys: An experiment with Web, mail, and yelephone. *Epidemiology*, *16*, pp. 701-709.

McCabe, S. E. (2004), "Comparison of Web and Mail Surveys in Collecting Illicit Drug Use Data: A Randomized Experiment", *Journal of Drug Education*, 34, pp. 61-72.

McCabe, S. E., C. J. Boyd, M. P. Couper, S Crawford, and H. D'Arcy (2002), "Mode Effects for Collecting Alcohol and Other Drug Use Data: Web and U.S. Mail", *Journal of Studies on Alcohol*, 63, pp. 755-761.

McCabe, S. E., M. P. Couper, J. A. Cranford, and C. J. Boyd (2006), "Comparison of Web and Mail Surveys for Studying Secondary Consequences Associated with Substance Abuse: Evidence for Minimal Mode Effects", *Addictive Behaviors*, *31*, pp. 162-168.

Schwarz, N., B. Knäuper, H.-J. Hippler, E. Noelle-Neumann, and F. Clark (1991), "Rating Scales: Numeric Values May Change the Meaning of Scale Labels", *Public Opinion Quarterly*, 55, pp. 618-630.

Tourangeau, R., F. G. Conrad, and M. P. Couper (2013), *The Science of Web Surveys*. New York: Oxford.

Tourangeau, R., M. P. Couper, and F. G. Conrad (2004), "Spacing, Position, and Order: Interpretive Heuristics for Visual Features of Survey Questions", *Public Opinion Quarterly*, 68, pp. 368-393.

Tourangeau, R., M. P. Couper, and F. G. Conrad (2007), "Color, labels and interpretive heuristics for response scales", *Public Opinion Quarterly*, *71*, pp. 91-112.

Tourangeau, R., M. P. Couper, and D. M. Steiger (2003), "Humanizing Self-Administered Surveys: Experiments on Social Presence in Web and IVR Surveys", *Computers in Human Behavior*, 19, pp. 1-24.

Tourangeau, R., and T. Yan (2009), "Sensitive Questions in Surveys." *Psychological Bulletin*, 133, pp. 859-883.