PICTURE THIS!
EXPLORING VISUAL EFFECTS IN WEB SURVEYS

MICK P. COUPER
ROGER TOURANGEAU
*University of Michigan and
Joint Program in Survey Methodology*
KRISTIN KENYON
Knowledge Networks, Inc.

Among the potential advantages of conducting self-administered surveys over the Internet are the rich visual possibilities offered by the Web. These varied visual resources include an assortment of typefaces, colors, and other stylistic elements; in addition, the Web offers the opportunity to enrich the survey experience with images, sound, and video. Perhaps the most promising of these visual enhancements is the ability to deliver color photographs or other images to respondents. Such images could help clarify items that assess brand recognition, magazine readership, or other topics in which a picture can help define the subject of interest.

The chance to extend survey measurement beyond the words of a question is one of the most exciting, yet least explored, aspects of Web surveys. It raises a host of related questions: What are the advantages and disadvantages of using images in a survey instrument? Do images convey more information than a researcher bargained for? Can well-chosen visual images clarify the meaning of the question for the respondent, or are images so intrinsically ambiguous that they generally confuse the meaning of the question (Mussio 1993, p. 325)? The authors of survey questions have long struggled to write clear survey questions—ones that are easy to understand and are understood by all respondents in the same way—and it remains to be seen whether the addition of images and other visual enhancements to survey questions will help achieve that goal or make it harder than ever to reach. This paper reports on some initial efforts to explore the role of images in the Web survey context.

We gratefully acknowledge the support of Knowledge Networks, Inc. in developing and deploying these surveys. We also thank the editor and anonymous reviewers for their helpful comments and suggestions. Address correspondence to Mick P. Couper; e-mail: mcouper@umich.edu.

68:255–266
DOI:10.1093/poq/nfh013

VISUAL AND VERBAL ELEMENTS OF SURVEY INSTRUMENTS

Following Ware (2000), we distinguish verbal or textual information, on the one hand, from visual information, on the other. In surveys the verbal information encompasses question wording, instructions, and certain auxiliary cues, such as the numbers used to label the scale points. The visual information includes standard features present in all questionnaires such as size and font of the type, color, position on the page, layout, and graphical symbols (such as arrows or boxes); it also encompasses such additional features as images, line drawings, and animation (cf. Redline and Dillman 2002).

Our study concerns the relationship between the verbal text of a question and the content of any accompanying images. The two kinds of information may be processed in parallel (e.g., see Paivio's [1986, 1991] dual coding model), allowing them to interact in complex ways. For example, if the verbal and visual information are inconsistent with each other, interference effects can occur. A well-known example of such interference is the Stroop effect. Subjects are asked to say what color a word is printed in. If the word is a color word and it names a color different from the print color, it slows subjects down (e.g., Deregowski, Parker, and McGeorge 1999; MacLeod 1991; Zimmer 1999). On the other hand, when the text and visual cues are consistent (e.g., the word "red" is printed in red letters), accentuation effects can occur (e.g., Krueger and Rothbart 1990). There are two other potential outcomes of the interaction between the verbal and visual features of a question. One possibility is that the visual information, intended as an embellishment to the text, draws attention away from the more important verbal information, producing distraction effects. A final possibility is that the visual information is the basis for unintended inferences that lead to misinterpretations of the verbal information. Schwarz and his colleagues have demonstrated similar misleading inferences in self-administered questionnaires (Schwarz 1994, 1996). For example, in one study, respondents answered differently when the scale end points were labeled +5 and -5 versus 0 and 10 (Schwarz et al. 1991; see also Schwarz, Grayson, and Knäuper 1998).

TASK VERSUS STYLISTIC ELEMENTS

These inferences in part reflect respondent uncertainty about the intended use of a feature of a question. Some features—"the task elements"—are essential to the task of completing the survey. These include the question wording, the response options, any instructions, navigational cues, and so on. These are typically verbal elements (including numeric elements, such as the question number and response scales), but they may include visual elements (arrows, boxes, radio buttons, etc.) as well, as Redline and Dillman (2002) argue. Moreover, in some cases (for example, brand recognition studies), the task elements could consist of images or pictures. In addition to the task elements, there are features of the instrument that are incidental or unrelated to the task itself. We refer to these as “style elements.” The style elements include the overall “look and feel” or layout of the Web site or survey instrument, the typeface used to represent the survey questions, background color, branding elements, and so on. Style elements are typically but not necessarily visual.

Although there is some overlap, the classification of the elements by their content is distinct from their classification by intended use. Task elements (survey questions) can be presented using either words or pictures but typically rely on the former. Similarly, the style elements can include words (e.g., contact information, survey organization name), pictures (e.g., logo), or both.[1]

As Schwarz’s studies show, although the survey designer may view an element as stylistic, respondents may see it as relevant to the tasks of understanding and answering the questions. There is ample evidence of such interpretive errors in both self-administered (Redline and Dillman 2002; Smith 1995) and interviewer-administered surveys (Couper and Hansen 2001; Sanchez 1992; Smith 1995). Just as incidental verbal features of the questions may lead to unintended inferences about the meaning of the questions (see Clark and Schober 1992; Schwarz 1994, 1996), incidental visual cues may produce unintended interpretive inferences as well.

Decisions about what the designers intend as style elements are, therefore, important in constructing a survey instrument. Some style decisions directly affect the ability of a respondent to complete the task. For example, the use of a fancy script font may render the question virtually unreadable for the respondent. Similarly, poorly chosen foreground/background color combinations may not produce enough contrast to make the question text easy to read. But there are other, subtler effects that can arise from style decisions, including color or size variations that steer respondents toward one response option rather than another, drop boxes that make some options less visible than others, and so on.

THE IMPACT OF IMAGES

Our study focuses on one particular visual element, namely, the use of photographic images to supplement the question text. The task-style continuum suggests several different uses of pictures in Web surveys. These include:

1. Questions in which images play an essential role (such as questions on recall of an advertisement, brand recognition questions, questions on magazine readership);
2. Questions in which images supplement the question text, whether the images are intended as motivational embellishments or as illustrations of the meaning of the question;

[1] Redline and Dillman’s (2002) notion of “auxiliary language”—visual elements that accompany text and contribute to communication—can thus be viewed as task-related visual elements.

3. Questions in which the images are incidental (providing branding, an attractive background, etc.).

Questions exhibiting all three methods of blending text and image appear to be quite widespread in Web surveys (see, for example, the Lightspeed Panel, online at http://us.lightspeedpanel.com). The arguments for questions using the first type of text-image combination are quite compelling, and questions in the third category—in which the images are incidental to the task—may also make sense in the highly competitive world of Web surveys, where branding is an important goal of many purveyors of Web surveys and services. Questions in which images are intended to play a supplementary role are potentially the most problematic because it may not be clear to respondents whether the images are intended as task or style elements. However, even though the use of images to supplement survey questions is common practice, we know of no evidence that provides a clear rationale for this type of question—no one has yet demonstrated that such visual embellishments improve respondent motivation, increase accuracy, reduce breakoffs, or decrease missing data.

We conducted a study to explore the relationship between the verbal and visual information in Web surveys; its focus is the impact of images on questions where the images present information designed to supplement the question text. Specifically, we explore how the content of the image may change the interpretation of the accompanying question for the respondent.

Sample

The experiment was run in a survey administered to a national sample of U.S. adults from the Knowledge Networks Panel. The panel is made up of approximately 100,000 panel members from almost 50,000 households in the United States. Recruitment into the panel begins with a list-assisted random digit dial (RDD) sample. When a household contacted by telephone agrees to participate in the panel, Knowledge Networks provides it with free hardware, a WebTV unit (an Internet device that connects to a television and a phone line), free Web access, password-protected e-mail accounts for each household member age 13 and older, ongoing technical support, and an incentive program to encourage continued participation in the panel. Each panel member receives the same Internet hardware and service, which help assure that each survey looks the same to every panel member. This feature is especially important when measuring the effects of visual images in surveys, as there is no need to adjust for differences in Web browsers or computer software and hardware. Of course, there may still be variation across respondents because of differences in the television screen size and resolution, the placement of the television in the house, and so on. The first survey assigned to panel members captures key demographic profile information. It is assigned to the adult members and is available as soon as the household connects the WebTV unit. The overall response rates for the recruitment efforts are as follows. Using the American Association for Public Opinion Research (2000) response rate definition RR2, 56 percent of contacted households agree to join the panel; 80 percent of those who agree to join actually install their WebTV unit; and 83 percent of those complete the core member and core household profile. Across these several stages, the overall response rate is approximately 37 percent. These are average estimates, as panel recruitment is an ongoing effort. See Krotki and Dennis (2001) and the panel's Web site (www.knowledgenetworks.com) for more details on the design and implementation of the panel.

The sample for our study was randomly selected from the Knowledge Networks profiled panel. The random sample was drawn from all available and active panel members who had not already been selected for other surveys being conducted by Knowledge Networks during the same weeks as our survey. Only one adult panel member per household was eligible for the survey. It concerned travel, leisure, and shopping activities and obtained an 80 percent completion rate, with 2,385 completed interviews from an initial assignment to 2,996 panel members. Taking into account the losses at earlier stages of recruitment and data collection, the cumulative response rate was 30 percent.

The analyses we present are based on unweighted data, as our focus is on the differences between experimental conditions rather than on national projections. Data from the Knowledge Networks panel are usually weighted to reflect selection probabilities and poststratified in an effort to compensate for noncoverage and nonresponse. The weighted analyses yield essentially the same results as those presented here. The sample respondents were primarily females (54.0 percent) and whites (81.7 percent), though a substantial number of African Americans (10.3 percent) completed the survey as well; 9.7 percent of the respondents described themselves as Hispanic. Most of the respondents were married (59.1 percent) or single (20.8 percent); the remaining 20.1 percent were widowed, separated, or divorced. The sample included a broad age range (from 18 to 94 years old), with a median age of 44. Finally, 17.0 percent of the sample had never finished high school, 25.2 percent were high school graduates, 35.9 percent had some college or an associate's degree, 14.5 percent were college graduates, and an additional 7.5 percent had advanced degrees.

Travel, Leisure, and Shopping Questionnaire

The questionnaire for the study focused on several common activities including travel, attending sporting events, listening to live or recorded music, dining out, and shopping. For each topic, we developed four versions of the questions: (1) a version that did not include any picture (the "no picture" condition); (2) a version featuring an image of a salient, but low frequency instance of the behavior in question (the "low frequency" condition); (3) a version featuring an image of a salient, high frequency instance (the "high frequency" condition); and (4) a version that displayed both pictures (the "both pictures" condition).[2] We randomly assigned respondents to one of the four conditions independently for each topic. Our hypothesis was that presenting the picture of the high frequency instance would enhance the retrieval of such instances and increase the total number of instances reported. By contrast, the picture of the low frequency instance would trigger the recall of relatively infrequent incidents similar to the one in the picture. For example, we asked respondents about their shopping trips in the past month and expected that showing them a picture of a grocery store would increase the overall number of shopping trips they reported on average compared to the picture of a department store, since trips to the grocery (cued by the one picture) are likely to be more frequent than trips to a department store (cued by the other).

For two of the topics in our study, we subsequently carried out a follow-up study to confirm that the pictures did in fact portray highly salient instances of the category. The follow-up study was part of a Web survey, in which we purchased a sample of e-mail addresses from Survey Sampling, Inc. (SSI); SSI sent e-mail messages inviting members of the sample to take part in the study. A total of 2,568 of them completed the questionnaire (out of 14,282 in the sample). The questionnaire included questions asking the respondents how often they went shopping and how often they took overnight trips. Just after the frequency question on shopping, respondents were asked "which of the following types of store did you consider in answering the previous question," with grocery stores and department stores among the possibilities listed. (Respondents were asked to pick all of the types of store they had considered.) A subsample answered an open-ended question instead: "When thinking about shopping, what's the first type or store or shop that comes to mind?" Similarly, we asked respondents "which of the following types of trips" they had had in mind in answering the prior question on their travel frequency. (Again, a subsample got an open-ended version of this question instead.) The open and closed results were similar for both shopping and travel, and we discuss only the responses to the closed items here. Grocery stores were the most commonly mentioned type of store, with 93.2 percent of the respondents indicating they had considered them in responding to the item about how often they went shopping Department stores were the next most popular choice (64.9 percent; another 5.9 percent mentioned clothing stores but not department stores). For the travel item the most popular choices were family vacations by car (76.9 percent), family visits by car (65.6 percent), and vacations by plane (50.2 percent). Business trips by plane were mentioned by 24.9 percent of the respondents. In the absence of any pictures, then, respondents were likely to consider these instances in assessing the frequency of shopping and

[2] The experimental conditions are presented in the appendix available in the online version of this journal.

Table 1. Images Displayed (and Sample Sizes), by Condition and Topic

	Picture Descriptions			
Question Topics	No Picture	Low Frequency Instance	High Frequency Instance	Both Pictures
Overnight trips in past year	(579)	Businessman at airport (620)	Family station wagon (593)	(593)
Sporting events attended in past year	(582)	Large baseball stadium (621)	Little League ball game (646)	(536)
Times went out to eat in past month	(592)	Intimate restaurant (593)	Eating fast food in a car (585)	(615)
Live music events attended in past year	(608)	Large outdoor concert (608)	Piano and singer at club (572)	(597)
Listening to recorded music in past week	(591)	Listening to the hi-fi (588)	Listening to the car radio (598)	(608)
Shopping trips in past month	(616)	Department store (clothing) (594)	Grocery store (548)	(627)

traveling—they are highly salient examples. Still, for some respondents, the pictures were likely to remind them of incidents they might otherwise have forgotten or overlooked. Table 1 summarizes the experimental design and sample sizes for each topic.

Results

The questions on each topic included an item asking how often the respondent engaged in the behavior, with the reference period tailored to the specific behavior. In addition, there were one or more follow-up questions on each topic. For example, we asked respondents who reported any overnight travel what proportion of these trips were for business. For each topic, we hypothesized that the picture showing a low frequency instance of the behavior (e.g., a businessman at an airport, a couple dining in a restaurant) would yield lower frequency reports than the picture showing a high frequency instance (a family in a station wagon, a woman eating fast food). The high frequency pictures would, we thought, bring more incidents to mind than the low frequency pictures. We also expected that including both pictures would further increase the reported frequency of engaging in the activity by reminding them of the full range of instances the category encompassed. We included the “no picture” condition as a control and expected that the frequencies in this condition would fall between those of the two main picture groups, since many respondents (but not all) would think of the high frequency instances anyway.

There were several outliers in the responses to each of these questions; we coded as missing all values that were three or more standard deviations from the mean. For example, for the overnight trips item, we dropped 31 respondents who reported more than 50 trips in the last year. For each item, we dropped the most extreme 1 or 2 percent of the observations. This deletion of outliers did not change the pattern of the means across the treatment groups for any of the six items, but it did sharply reduce the standard errors. The resulting means for each treatment for the six items is shown in table 2.

We compared the four means for each topic using one-way analysis of variance tests (ANOVAs). For all six topics the overall *F*-tests were significant

Table 2. Mean Numbers of Events Reported (and Standard Errors), by Condition

	Picture Descriptions				
Question Topics	Low Frequency Instance	No Picture	High Frequency Instance	Both Pictures	Significance Tests
Overnight trips in past year	4.44 (.25)	5.44 (.35)	5.73 (.33)	5.32 (.29)	*F* = 3.35, *df* = 3,2335, *p* = .018
Sporting events attended in past year	2.66 (.24)	3.12 (.26)	3.65 (.29)	3.42 (.30)	*F* = 2.55, *df* = 3,2299, *p* = .054
Times went out to eat in past month	9.86 (.44)	11.96 (.51)	13.64 (.55)	12.18 (.48)	*F* = 9.67, *df* = 3,2310, *p* < .0001
Live music events attended in past year	1.81 (.16)	1.50 (.14)	1.86 (.17)	2.22 (.19)	*F* = 3.28, *df* = 3,2342, *p* = .020
Listening to recorded music in past week	10.56 (.65)	10.71 (.63)	10.15 (.62)	12.52 (.70)	*F* = 2.67, *df* = 3,2273, *p* = .046
Shopping trips in past month	7.73 (.34)	8.72 (.36)	9.07 (.36)	9.79 (.37)	*F* = 5.81, *df* = 3,2313, *p* = .0006

(p = .05 or less). In addition, for four of the six topics the means for the high and low frequency conditions differed significantly from each other ($p < .01$); the two exceptions involved live music and recorded music. In all four cases the difference is in the expected direction, with the picture showing the high frequency instance of the behavior prompting higher reporting on the average than the picture showing the low frequency instance.

The effects of the images were also apparent in several of the follow-up questions to the key behavioral frequency items. For example, the question on sporting events was followed by one asking when the most recent event had taken place; 29.6 percent of those in the condition showing a major league ballpark responded “more than 6 months ago” versus 20.2 percent of those in the Little League game condition ($\chi^2 = 7.02$, $df = 1$, $p = 0.0081$). Similarly, those who got the intimate restaurant picture with the dining out question were significantly ($p < .05$) more likely to say they enjoyed their last meal very much (44.3 percent) than those who got the fast food picture (37.7 percent); the former group also reported spending a significantly ($p < .01$) higher amount of money on the typical meal outside the home ($14.46 versus $12.37; $F = 15.41$, $df = 1{,}1087$, $p < .0001$).

We looked for interactions between the experimental treatment and the age, sex, and education of the respondents and found no consistent patterns; only 2 of 18 interaction effects we examined were significant. We predicted that the “no picture” condition would produce reported frequencies between those for the two main picture conditions; again with the exception of listening to live and recorded music, we find this to be the case. Contrary to our prediction, however, the condition that included both images did not consistently yield the highest levels of reporting. In general, the results suggest that providing both images yields levels of reporting similar to that of the high frequency condition. The high frequency picture may have triggered the retrieval of similar numbers of incidents regardless of whether it was presented alone or accompanied by a less frequent instance of the category.

In summary, the results of our study, in which the questions involved familiar categories and the relevance of the images was clear, suggest that the use of specific images to supplement the text of survey questions can affect the responses provided in systematic and predictable ways.

Discussion

A key argument for including pictures or other visual enhancements in Web surveys is to increase respondents’ enjoyment of or interest in the survey, thereby reducing the frequency of breakoffs. There are low rates of breaking off in the Knowledge Networks panel (once panel members start a survey, they almost always complete it), but we can examine the validity of this argument by looking at responses to two debriefing items included at the end of our survey. One asked respondents to rate their satisfaction with the survey; the other asked them to rate the survey's length. To measure the effect of including visual images, we created a simple index; respondents who were exposed to a given picture were coded 1 and those who were not were coded 0. We summed across the six experiments so that the index ranged from 0 (no pictures seen) to 6 (pictures seen in all six experiments). The number of pictures respondents saw did not relate to either their ratings of their satisfaction or the length of the survey (r = –0.012 for satisfaction and r = 0.031 for length). Thus, we find little support for the argument that including images increases respondents' enjoyment of the interview or reduces the perceived burden.

The results do suggest that the use of supplemental images can systematically influence answers to survey questions. Our experiments focused on one particular use of images—as illustrative supplements to the text of the survey questions. Although we cannot say whether the inclusion of photographs increased or decreased the accuracy of reporting, the content of the images in the study definitely affected the answers. Responses to an open-ended debriefing question at the end of our survey suggested that the pictures may not only have primed specific memories, but also affected how respondents construed the category. This was most noticeable for the question on shopping frequency, which was followed by a question on the proportion of shopping trips that were for food. Several respondents commented on the impact of the images, for example:

> "What kind of shopping you were looking for was not defined because my number of times would be different depending on what type. I took it as how many times for leisure" [no picture].
>
> "Thought shopping meant clothes from picture. If you include food shopping—went about 10 times" [department store picture].
>
> "I shop for groceries almost every week. Does that count? The pictures are nice, but add to the time it takes to answer a survey" [department store picture].
>
> "The pictures helped remind me that a little league game is just as much a sporting event as a trip to Fenway. The pics were a help" [both sporting event pictures].

For some respondents, the pictures clarified the meaning of the questions, broadening their definition of the target category. For others, the pictures may have reinforced a narrow interpretation of the question's meaning.

Given that we find no evidence that images boost respondents' motivation to complete surveys, we suggest caution in adding such visual embellishments to Web questionnaires. Pictures convey rich information and can trigger the recall of some incidents, though perhaps at the expense of others. Pictures can also suggest either a broader or narrower construction of the category of interest than the text of the question would otherwise convey. The relative weight that respondents give to the words and pictures may also affect the responses provided. It is possible that the pictures in our experiments had more impact because the questions concerned relatively broad, poorly defined categories like shopping. If the target categories had been narrower (e.g., grocery shopping), the pictures might have affected the answers less.

Clearly, this research is only a preliminary foray into the issue of how images affect responses to survey questions. The results from our study suggest some fruitful lines of future inquiry. But one thing is clear already—if pictures are indeed worth a thousand words, we should be sure to choose those pictures with great care.

References

American Association for Public Opinion Research (AAPOR). 2000. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. Lenexa, KS: AAPOR.

Clark, Herbert H., and Michael F. Schober. 1992. “Asking Questions and Influencing Answers.” In *Questions about Questions: Inquiries into the Cognitive Bases of Surveys*, ed. Judith M. Tanur, pp. 15–48. New York: Russell Sage Foundation.

Couper, Mick P., and Sue Ellen Hansen. 2001. “Computer Assisted Interviewing.” In *Handbook of Interviewing*, ed. Jaber F. Gubrium and James A. Holstein, pp. 557–75. Thousand Oaks, CA: Sage.

Deregowski, J. B., D. M Parker, and P. McGeorge. 1999. “What’s in a Name? What’s in a Place? The Role of Verbal Labels in Distinct Cognitive Tasks.” *Current Psychology* 18(1):32–46.

Krotki, Karol P., and J. Michael Dennis. 2001. “Probability-Based Survey Research on the Internet.” Proceedings of the 53rd Conference of the International Statistical Institute, Seoul, Korea.

Krueger, Joachim, and Myron Rothbart. 1990. “Contrast and Accentuation Effects on Category Learning.” *Journal of Personality and Social Psychology* 59(4):651–64.

MacLeod, C. M. 1991. “Half a Century of Research on the Stroop Effect: An Integrative Review.” *Psychological Bulletin* 109(2):163–203.

Mussio, P. 1993. “Representational Problems in Visual Language Design.” *Journal of Visual Languages and Computing* 4:325–26.

Paivio, Allan. 1986. *Mental Representations: A Dual-coding Approach*. New York: Oxford University Press.

———. 1991. “Dual Coding Theory—Retrospect and Current Status.” *Canadian Journal of Psychology* 45:255–87.

Redline, Cleo D., and Don A. Dillman. 2002. “The Influence of Alternative Visual Designs on Respondents’ Performance with Branching Instructions in Self-Administered Questionnaires.” In *Survey Nonresponse*, ed. Robert M. Groves, Don A. Dillman, John L. Eltinge, and Roderick J. A. Little, pp. 179–93. New York: Wiley.

Sanchez, Maria E. 1992. “Effect of Questionnaire Design on the Quality of Survey Data.” *Public Opinion Quarterly* 56:206–17.

Schwarz, Norbert. 1994. “Judgment in a Social Context: Biases, Shortcomings, and the Logic of Conversation.” *Advances in Experimental Social Psychology* 26:123–62.

———. 1996. *Cognition and Communication: Judgmental Biases, Research Methods, and the Logic of Conversation*. Mahwah, NJ: Lawrence Erlbaum Associates.

Schwarz, Norbert, Carla E. Grayson, and Bärbel Knäuper. 1998. “Formal Features of Rating Scales and the Interpretation of Question Meaning.” *International Journal of Public Opinion Research* 10(2):177–83.

Schwarz, Norbert, Bärbel Knäuper, Hans-Jurgen Hippler, Elisabeth Noelle-Neumann, and Leslie Clark. 1991. “Rating Scales: Numeric Values May Change the Meaning of Scale Labels.” *Public Opinion Quarterly* 55:618–30.

Smith, Tom W. 1995. “Little Things Matter: A Sampler of How Differences in Questionnaire Format Can Affect Survey Responses.” *Proceedings of the American Statistical Association, Survey Research Methods Section*, pp. 1046–51.

Ware, Colin. 2000. *Information Visualization: Perception for Design*. San Francisco: Morgan Kaufman.

Zimmer, Hubert D. 1999. “Inappropriate Colors Impair Word-Picture and Picture-Word Matching.” *Current Psychology of Cognition* 18(1):3–25.