Session 3

The Impact of the Visible:

Images, Spacing, and Other Visual Cues in Web Surveys

The Impact of the Visible:
Images, Spacing, and Other Visual Cues in Web Surveys

Roger Tourangeau, Mick P. Couper, and Fred Conrad
Survey Research Center, University of Michigan
Joint Program in Survey Methodology, University of Maryland

Introduction

The rapid acceptance of the Web as a vehicle for survey data collection raises important questions for survey designers. Web surveys are the latest example of computerized self-administration of survey questions, and we suspect they may ultimately turn out to be the most popular. Aside from the gains from computerization and self-administration, Web data collection eliminates interviewers entirely, sharply reducing the cost of data collection. Furthermore, Web surveys can deliver rich visual content that is impossible or prohibitively expensive to incorporate in other modes. Not surprisingly, the growth in Web surveys has been dramatic. Despite serious concerns about coverage and nonresponse in Web surveys (Couper, 2001), the commercial research sector has rapidly embraced the Internet for faster and cheaper data collection, and almost daily there are reports of new surveys being done over the Web.

A key characteristic of Web surveys is their reliance on visual presentation of the questions. Of course, sound can be added to Web questionnaires, but so far Web surveys have remained a visual medium. Visual presentation is not unique to Web data collection, but is shared to varying degrees with most other methods of self-administration, including mail surveys.

Still, the implications of visual presentation are not especially well understood, even for the older methods; the literature on the design of mail or paper-based self-administered questionnaires is not large. Although several good texts offer practical guidelines for the design of paper self-administered questionnaires (e.g., Dillman, 1978; Mangione, 1995), there has been relatively little empirical work or theoretical analysis of the issues involved. The forms design literature is sparse in general (see, e.g., Burgess, 1984; Waller, 1984; Wright and Barnard, 1975). The one notable exception has been the work of Redline and Dillman, who have applied principles rooted in visual perception theory to the design of self-administered forms (Dillman, Redline, and Carley-Baxter, 1999; Jenkins and Dillman, 1995; Redline and Dillman, 2002). The focus of this work has been on designing forms so that respondents are willing and able to complete them. But the design of paper forms and computer screens may affect not only whether respondents answer the questions but also which answers they give (e.g., Sanchez, 1992; Smith, 1995). The study of forms design is in its infancy, and the impact of forms design on measurement error has been almost entirely neglected.

The studies we present here support a few general conclusions about the impact of visual information on responses to questions in Web surveys:

- Respondents notice images in Web surveys and the content of these images can affect the answers they give;

- Respondents also take in such visual cues as the spacing and relative position of the response options and these cues can alter their interpretation of survey questions;

- Respondents are sensitive to information that is immediately visible and may ignore information that is equally critical but not equally available.

Taken together, our results suggest that, whether we want them to or not, respondents attend to the visual design of Web questionnaires as well as to the verbal content of the questions.

Images as Context

One line of our work has focused on the use of photographic images to supplement question text. As we have argued in an earlier paper (Couper, Tourangeau, and Kenyon, in press), visual and verbal elements may be essential to complete the task of understanding and responding to the questions or these elements may be inessential stylistic embellishments that create an overall "look-and-feel" for the questionnaire. This task-style continuum suggests several different ways pictures can be used in Web surveys:

1. Questions in which images play an essential role (such as questions on recall of an advertisement, brand recognition questions, questions on magazine readership, etc.);

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;

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

All three combinations of text and image appear to be quite widespread in Web surveys. 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 elements or style elements.

Whether the survey designers intend it or not, images can serve as powerful contextual cues that alter what material comes to mind as respondents formulate their answers; they can affect how respondents construe the targets of their judgments or the standards they apply in making those judgments. Let us briefly summarize the results of three studies that illustrate these processes.

Images, target categories, and frequency judgments. Our first experiments on the impact of images were done by Knowledge Networks, which embedded them in a survey administered to a 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, initially recruited from a list-assisted RDD sample. Each panel member receives the same WebTV unit and software, which help assure that the survey looks the same to every panel member. Some 56% of contacted households agree to join the panel but only 80% of those actually install the WebTV unit and only 83% of those complete the initial questionnaire which gathers basic demographic data on panel members. These are average estimates, as panel recruitment is an ongoing effort. Dennis (2001) provides more details on the design and implementation of the panel. About 3,000 members of the panel were asked to complete our survey, and 2,385 of them did. Taking into account the losses at earlier stages of recruitment and data collection, the cumulative response rate for our survey was 30%.

The survey, which concerned travel, leisure, and shopping activities, included six parallel experiments summarized in Table 1 below. All six followed the same logic. 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).

Our hypothesis was that presenting the picture of the high frequency instance would enhance the retrieval of similar 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).1

1For two of the topics in our study, we carried out a follow-up study to confirm that the pictures did in fact portray highly salient instances of the category. The follow-up 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.) Similarly, we asked respondents "which of the following types of trips" they had in mind in answering the prior question on their travel frequency. Grocery stores were the most commonly mentioned type of store, with 93.2% 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%; another 5.9% mentioned clothing stores but not department stores). For the travel item, the most popular choices were family vacations by car (76.9%), family visits by car (65.6%), and vacations by plane (50.2%). Business trips by plane were

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

Question topics	**Picture Descriptions**			
	No Picture	Low Frequency Instance	High Frequency Instance	Both Pictures
Overnight trips in last year	(579)	Businessman at airport (620)	Family station wagon (593)	(593)
Sporting events attended in last 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 the last year	(608)	Large outdoor rock concert (608)	Piano and singer at club (572)	(597)
Listening to recorded music in the past week	(591)	Listening to the hi-fi (588)	Listening to the car radio (598)	(608)
Shopping trips in the past month	(616)	Department store (clothing) (594)	Grocery store (548)	(627)

We compared the four means for each topic using one-way ANOVAs. For all six topics, the overall *F*-tests were significant. In addition, for four of the six topics, the means for the high and low frequency conditions differed significantly from each other (at $p < .01$ or less); the two exceptions involved live music and recorded music. In all four cases, the difference was in the expected direction, with the pictures showing the high frequency instances of the behaviors prompting higher reporting on the average than the pictures showing the low frequency instances. We interpret this as the same sort of accessibility-based context effects that are often found in attitude surveys (see Chapter 7 in Tourangeau, Rips, and Rasinski, 2000) — the images affect the number and type of instances respondents retrieve in formulating their answers ("priming" those memories); the number and type of instances retrieved in turn affect the judged frequency of the behavior.

Responses to an open-ended debriefing question at the end of our second survey suggested that the pictures may not only have primed specific memories, but also affected how respondents construed the category of interest. This was most noticeable for the question on shopping frequency, which

---

mentioned by 24.9% of the respondents. In the absence of any pictures, then, respondents were likely to consider these instances in assessing the frequency of shopping and 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.

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 relatively narrow interpretation of the question's meaning.

Images and rated health. In our initial studies, then, respondents exhibited what are sometimes called *assimilation* effects in the context effects literature. When they saw images of high frequency events, they reported higher frequencies; when they saw images of low frequency events, they reported lower frequencies. Verbal context (in the form of prior items) can sometimes have the opposite effect on answers to subsequent questions. When the prior questions suggest an extreme standard of comparison that respondents apply in judging later items, the target judgments are pushed in the direction opposite of the standard. For example, respondents may report liking a politician less when they rate an extremely popular politician first (Schwarz and Bless, 1992). We thought we could create similar judgmental contrast effects using images rather than prior questions to set the standard for the target items.

This experiment was embedded in a Web survey conducted by MSInteractive. In March and April of this year, MSInteractive sent e-mail invitations to 39,217 members of SSI's Web survey frame. The e-mail invitation asked them to complete a survey of attitudes and lifestyles sponsored by the National Science Foundation; it included the URL for the questionnaire. The SSI frame consists of some seven million e-mail addresses collected at various Web sites. A total of 3,179 persons started the questionnaire, 2,722 of them getting all the way through it. The response rate was 6.9 percent (not counting the partials) or 8.1 percent (counting them).

The experiment compared the impact of two pictures on respondents' judgments of their overall health (that is, responses to an item asking, "How would you rate your health?"). One group of respondents saw photograph of a healthy young woman jogging; another group saw a picture of a woman in a hospital bed. The experiment also compared three different positions for the picture — on the prior screen just before the health item, on the same screen in the survey header, or just to the left of the question text. Figure 1 displays examples of the pictures we used.

Figure 1. Images used in Study 2

a. Sick Woman — Picture in Header

Questions about this survey?
Email us at life@msiresearch.com
or call toll free 1.866.674.3375

How would you rate your health?

Extremely good	Good	Neutral	Poor	Extremely poor
○	○	○	○	○

Next Screen Previous Screen

b. Fit Woman — Picture to Left of Question

Questions about this survey?
Email us at life@msiresearch.com
or call toll free 1.866.674.3375

How would you rate your health?

Extremely good	Good	Neutral	Poor	Extremely poor
○	○	○	○	○

Next Screen Previous Screen

c. Fit Woman — Picture on Prior Screen

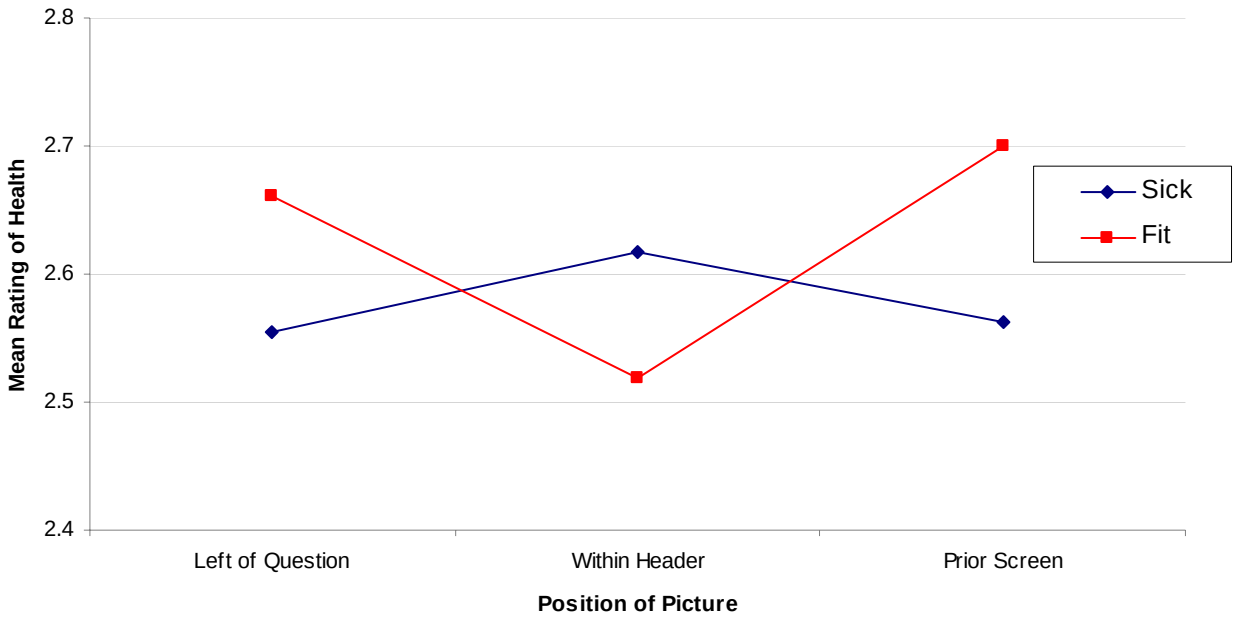
Questions about this survey?
Email us at life@msiresearch.com
or call toll free 1.866.674.3375

These first few questions ask about your health.

Next Screen Previous Screen

As expected, the pictures affected the self-ratings of health, lowering them on average for the respondents who got the picture of the healthy woman jogging (mean of 2.64) and raising them on average for those who got the picture of the sick woman in bed (2.58). (Higher numbers indicate worse health.) The overall effect of the picture was only marginally significant — $F(1, 2309) = 3.08$, $p < .08$. But we didn't expect the significant interaction between the position of the picture and its content; that interaction is displayed in Figure 2. When the picture is in the in the header, assimilation rather than contrast seems to be the result.

Figure 2. Health Ratings, by Picture and Position



At least in some conditions, then, images provide a standard of comparison against which our judgments of later targets, such as our own health, are contrasted.

Images in the interface. Tourangeau, Couper, and Steiger (2003) reported another series of experiments that incorporated images as part of the interface of a Web survey. (These studies were done by the Gallup Organization. Because these studies have been published, we omit the methodological details here.) Figure 3 below shows an example of the images the interface incorporated. The opening screen displayed the female face of one of the investigators (Steiger); other versions of the questionnaire displayed a male picture. Across two separate Web surveys, we examined the impact of the interface on answers to a variety of questions. For the most part, it didn't matter whether the survey had a male or a female "face," but for one set of items it did. These were a battery of questions on sex roles that are known to be affected by the sex of the (live) interviewer (Kane and Macauley, 1993). Men and women both give more pro-feminist responses to these items when female interviewers administer them than when male interviewers do. We found a similar pattern with our "virtual" interviewers; the responses were more pro-feminist when the

survey had a female "face" (as in Figure 3) than a male one. We suspect that this is a priming effect; when the respondents see the picture of an attractive working woman, it tends to bring to mind consistent (that is to say, positive) thoughts about women in the work place. The male interface tends to bring to mind more traditional views about the roles of men and women.

Figure 3. "Female" Interface used in Web Surveys

Researchers in human-computer interaction tradition have reported even more striking results. For example, Walker, Sproull, and Subramani (1994) administered questionnaires to people using either a text display or one of two talking-face displays to ask the questions. Those interacting with a talking-face display spent more time, made fewer mistakes, and wrote more comments than did people interacting with the text display. However, people who interacted with an expressive face liked the face and the experience less than those who interacted with an inexpressive face. In another experiment, Sproull and colleagues (1996) varied the expression of a talking face on a computer-administered career counseling interview; one face was stern, the other pleasant. The faces were computer-generated images with animated mouths. They found that:

> People respond to a talking-face display differently than to a text display. They attribute some personality attributes to the faces differently than to a text display. They report themselves to be more aroused (less relaxed, less confident). They present themselves in a more positive light to the talking-face displays. (p. 116).

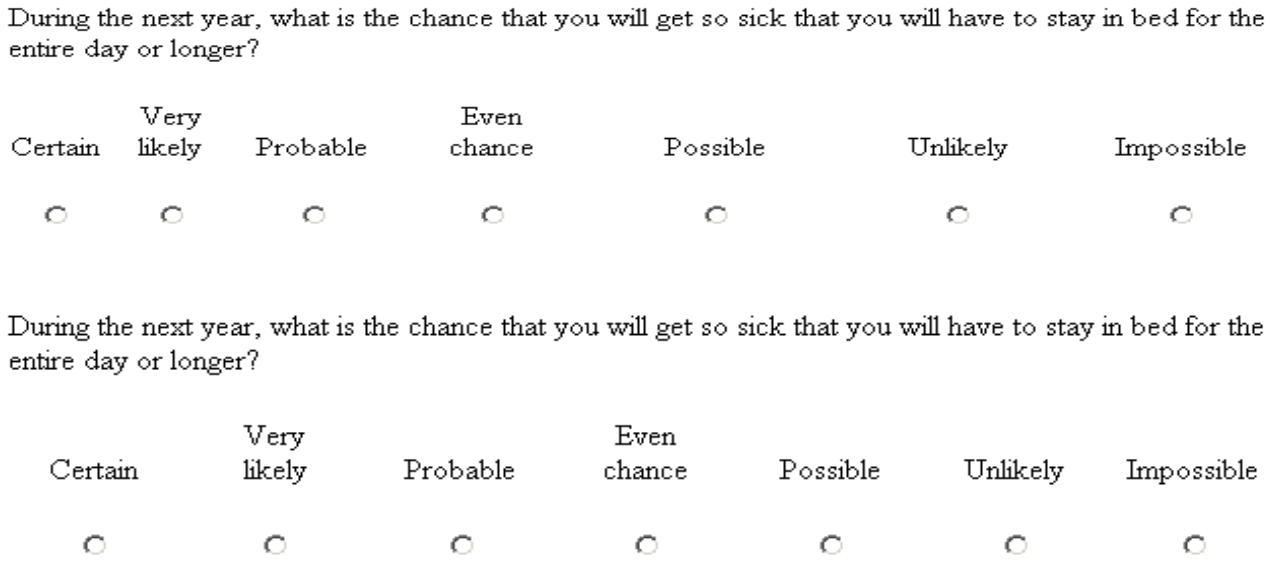
The interface to a survey, particularly when it incorporates humanizing visual cues, may itself constitute a contextual stimulus, one that is capable of altering respondents' views of the survey and their responses to the questions.

Spacing and Position

Spacing of the response options. Our experiment on self-rated health investigated a second issue besides the effect of the photographs. The question following the standard health item asked respondents how likely it was they'd get sick enough during the next year that they have to spend a day or more in bed ("During the next year, what is the chance that you will get so sick that you will have to stay in bed for the entire day or longer?"). We varied the spacing of the response options that made up the scale on which respondents were to indicate their answers. This experiment is one of a number we've done that share the notion that respondents follow simple heuristics in interpreting the visual features of questions. Though these interpretive heuristics are often useful, they may sometimes lead to unintended inferences about the meaning of a question. Hoffman (2000) argues that interpretive rules are central in visual processing and are responsible for such key abilities as depth perception. The heuristics for interpreting visual stimuli can sometimes lead to systematic misinterpretations of those stimuli, producing optical illusions. In the same way, the application of interpretive heuristics for visual cues in questionnaires can lead to erroneous inferences about the meaning of survey questions.

One of these heuristics involves seeing the option that is physically in the middle of the scale as representing the scale midpoint; we refer to this as the "middle means typical" heuristic. We varied the spacing of the response options to the question about the chance of a sick day in bed. Approximately half of the respondents got the item with evenly spaced response options (see Figure 4); the remainder got a scale in which four of the seven options were to the left of the visual midpoint of the scale.

Figure 4. Scales Used in Experiment on Spacing of Response Options

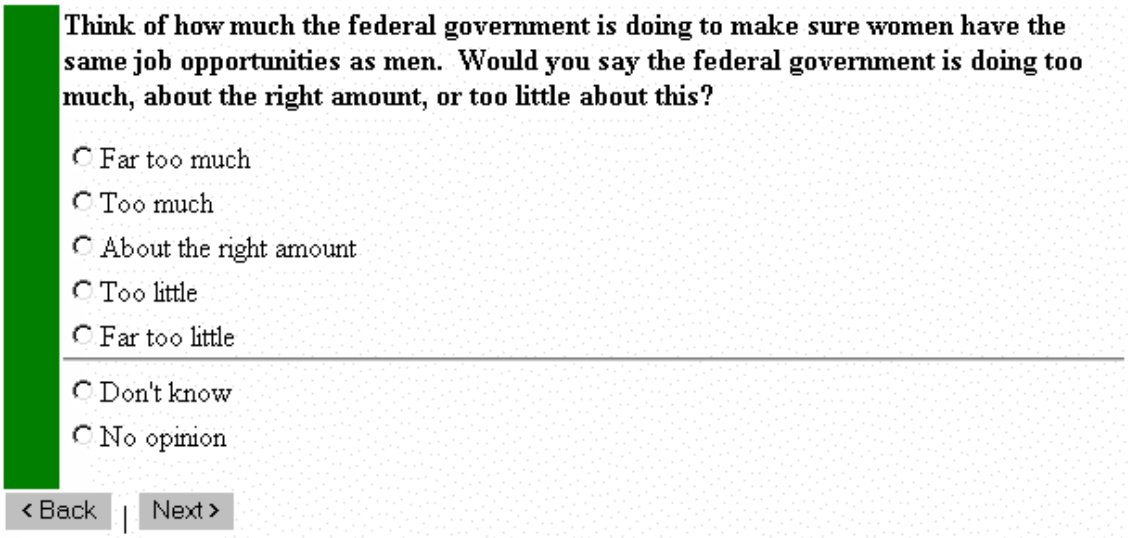


The ratings were significantly higher when respondents got the unevenly spaced scale (the top one in Figure 4) rather than scale that arrayed the response categories evenly (the bottom one). The means were 4.60 in the even spacing condition versus 4.45 in the uneven spacing condition; $F(1, 3083) = 7.58$, $p < .01$.

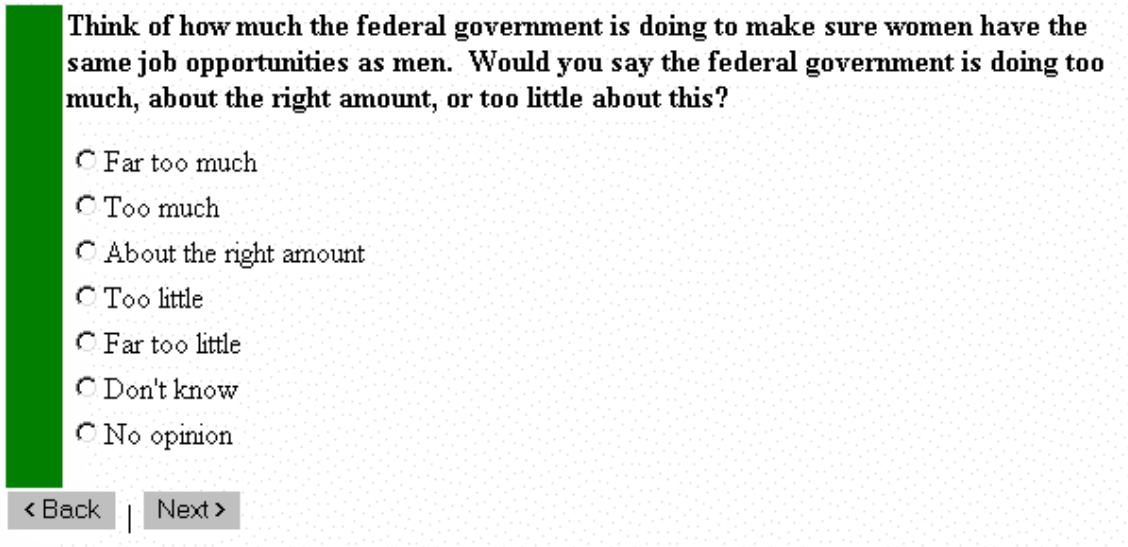
Separating substantive and nonsubstantive options. In one of our Gallup surveys, we did another study that demonstrated the importance of the spacing of the response options. That experiment compared two methods of separating nonsubstantive response options (Don't know, Refused) from substantive ones. In one case, the nonsubstantive options were simply presented as additional radio buttons; in the other, we included a divider line that clearly separated the nonsubstantive options from the rest (see Figure 5).

Figure 5. Formats for Displaying Nonsubstative Options

a. Divider Line Version



b. Version with No Divider Line



In the version with the divider line, the visual midpoint of the scale falls at the conceptual midpoint ("About the right amount"). In the version without the divider, the visual midpoint actually falls on one end of the scale ("Too little"). This difference affected the average responses — there's a significantly lower mean without the divider than there is with it. Moreover, the divider line seemed to draw attention to the nonsubstantive options; there are significantly more nonsubstantive answers given when the divider line is displayed (21.4%) than when it's omitted (17.5%).

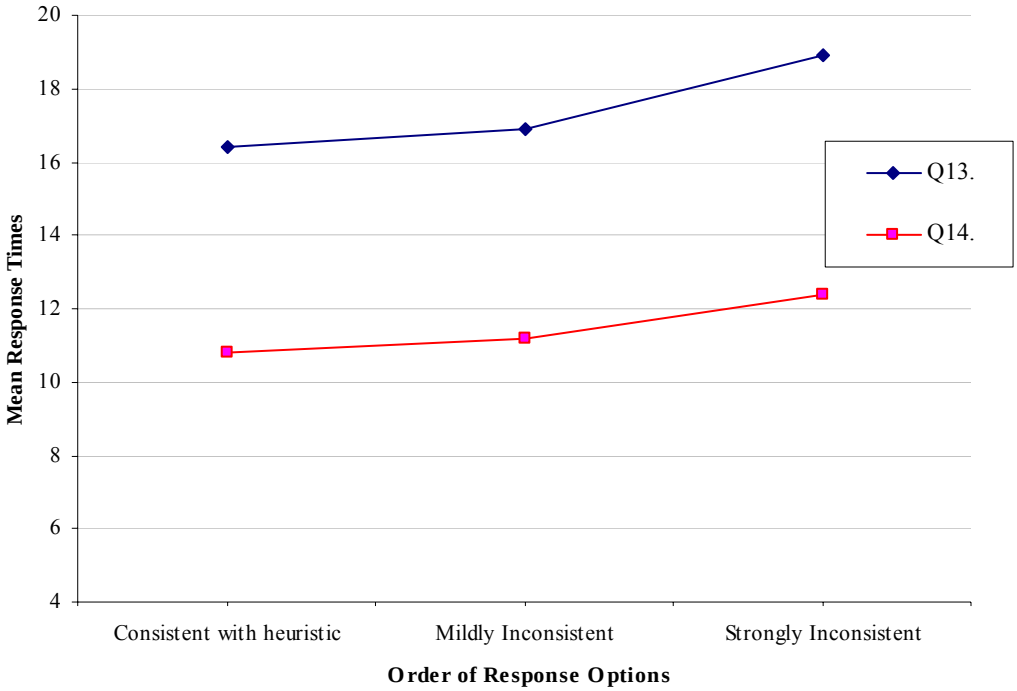
Positional inferences. Another heuristic respondents may use in understanding and applying response options is the "Left and top mean first" heuristic. According to the heuristic, the leftmost or top item in a list of items represents the "first" in some conceptual sense. For example, when the list is a series of ordered response categories or scale values, respondents expect the top or leftmost option to represent one of the two endpoints ("Agree strongly") and they expect each of the successive options to follow in some logical order ("Agree," "Neither agree nor disagree," and so on). If the list does not conform to these expectations, respondents may become confused, make mistakes, or take longer to respond.

Our first Web study with MSInteractive experimentally varied the order of the response options in six of the survey questions.2 We carried out two independent experiments, one with four frequency items and the other with two agree-disagree items. We focus here on the results from the agree-disagree items. Each experiment compared three versions of the questions. In one version, the response options followed the logical order. For the agree-disagree items, this version went from "Strongly agree" to "Strongly disagree" with "It depends" in the middle. A second version presented the options in order of decreasing agreement, with "It depends" as the final option). In the final version, "It depends" was the first option presented, but the remaining options were ordered by extremity ("Agree strongly," "Disagree strongly," "Agree," and "Disagree"). Respondents got all the response options in the same order for all four of the frequency items; similarly, they randomly assigned to receive one of the three versions for both agree-disagree items.

We anticipated that respondents would answer the questions most quickly when the items followed the order implied by the "left is first" heuristic, with the slowest answers in the third version of the questions (where the order of the response categories departs most sharply from the order implied by the heuristic). Three of the six items showed significant differences in response times and all three show the expected pattern. Figure 6 below displays the average response times for the two agree-disagree items — Q13 ("It is SENSIBLE to do exactly what the doctors say") and Q14 ("I have to be VERY ILL before I go to the doctor") in the experiment. For both items the differences in reaction times across experimental treatments were highly significant: $F(2,2533)=18.7$ for Q13 and $F(2,2591)=12.6$ for Q14.

2In February and March of 2002, MSInteractive conducted a Web survey, in which 14,192 e-mail invitations were sent to members of SSI's Web survey frame. The e-mail invitation asked them to complete a survey sponsored by the National Science Foundation. A total of 2,871 persons started the questionnaire, 2,568 of them getting all the way through it, for a response rate of 18.1% (not counting the partials) or 20.2% (counting them). Among other experiments, the survey included one that compared three response formats.

Figure 6. Response Times and Consistency with Heuristic



Another implication of the "Left is first" heuristic is that respondents may use it to infer the characteristics of an unfamiliar item from its position in a list of similar items. For example, we compared the percentage of respondents rating the Fiat Tipo as an expensive car when it came third in a list of cars that included the BMW 318, Acura Integra, Mazda Protégé, Toyota Corolla, Dodge Neon, and Geo Metro to when it came last seventh — right after the Geo Metro. Respondents were significantly more likely to say the Tipo was an expensive car when it came third in the list (72.4%) than when it came last (60.3%; $\chi^2 = 45.3$, df =1, $p < .001$). We found similar results for three out of five other items (see Table 2 below).

Table 2. Proportion Yes (and Sample Size), by Position in List

Judgment/Item	**Percent Yes** (n)	
	Third in List	Seventh in List
Important for healthy diet/Isoflavin	44.4 (1396)	43.2 (1326)
Low in saturated fat/Cod liver oil	42.7 (1396)	36.7 (1326)
Expensive hotel/Clarion Inn	61.5 (1396)	44.3 (1326)
Expensive city/Ocala,FL	51.5 (1396)	59.1 (1326)
Expensive midsize/Austin Rover	92.0 (1396)	86.0 (1326)
Expensive small car/Fiat Tipo	72.4 (1396)	60.3 (1326)

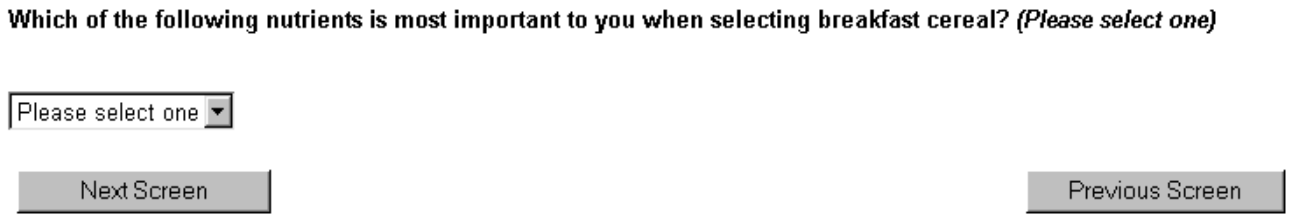
Note: The differences between columns are significant ($p < .05$ or less) for all but the first row.

The Visibility Principle

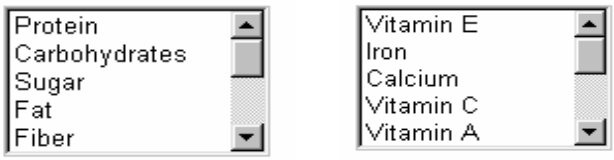
Images, spacing, and positioning all have an impact on the answers. We have also looked at how the method of presenting the response options can affect the distribution of the answers. We compared radio buttons (a format that displayed all 11 answer options from the outset), drop down boxes in which only the first five answer options were visible initially, and drop down boxes in which none of the options were visible until respondents clicked on the drop down arrow. Figure 7 shows the key conditions. In addition, the experiment varied the order of the answer options. Approximately half of the respondents got the response options in one order; the rest got them in the reverse order.

Figure 7. Formats Compared in Response Format Study

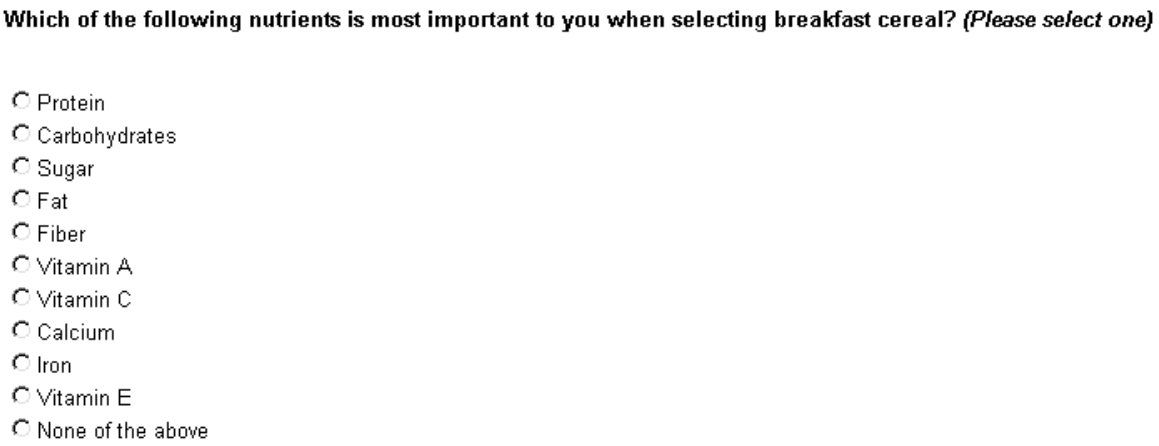
a. Drop Box — None of the Options Visible



b. Drop Box — Five Options Visible

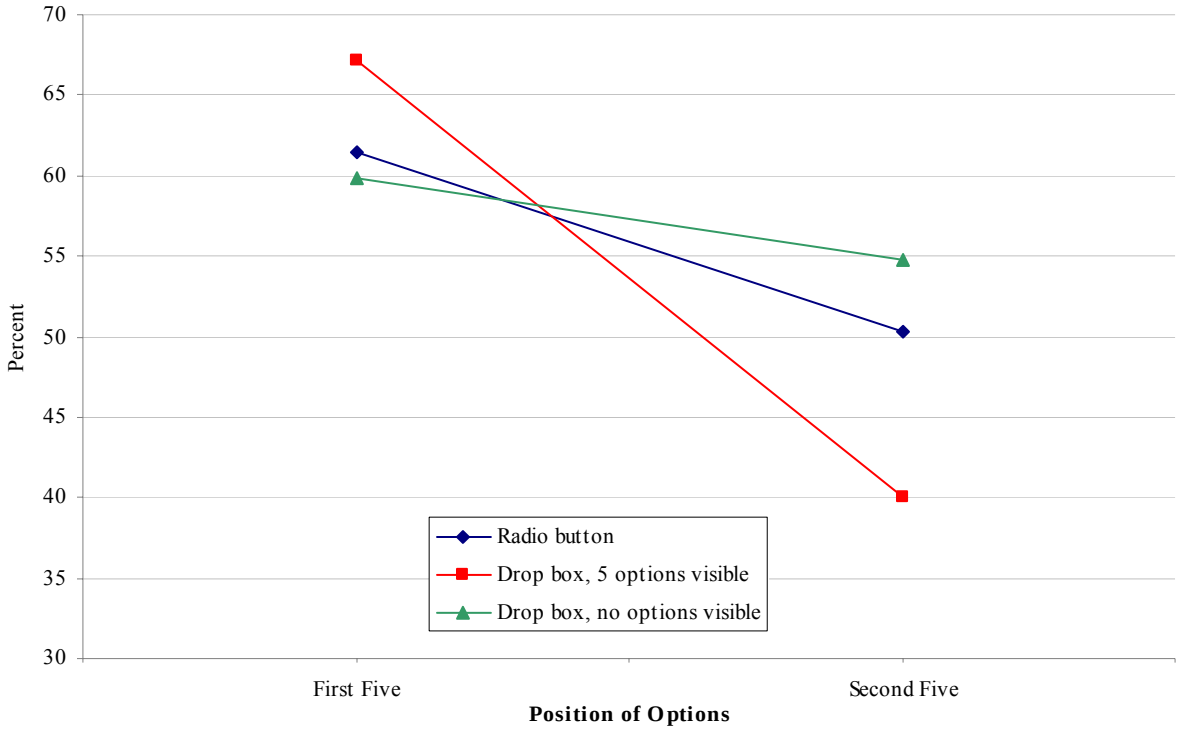


c. Radio Buttons



Our main hypothesis was the respondents would focus on the options they could see and thus would be more likely to select one of the five options displayed initially in the drop down box. Figure 8 displays the key result. In all three response formats, respondents were more likely to select one of the first five options listed in Figure 7c. above ("Protein" through "Fiber") when these were the first five options listed than when they were the final five. This is a classic primacy effect and similar effects are often found with items displayed visually (Krosnick & Alwin, 1987; Krosnick, 1991); however, the effect is far more marked in the drop down box condition in which only the first five options were visible. We included a replication of this experiment (Q20) later in the questionnaire, and the results are similar to those in Figure 8. The key interaction of response order and response format was highly significant for both items. When it takes additional effort to see some of the options, respondents are especially unlikely to choose them. Redline and Dillman (2002; see also Jenkins and Dillman, 1995) have also emphasized the importance of visual prominence.

Figure 8. Impact of Response Format and Response Order



Getting help. Our first MSInteractive survey also included experiments that examined the impact of the accessibility and usefulness of on-line definitions for survey terms. Respondents were asked to evaluate (on a five-point scale) whether they consume as much as they should of four food/nutrition products. They were told that they could obtain a definition for any of the terms by clicking on them but were not specifically instructed to do so. The primary concern in the study was how often they obtained definitions. Our initial hypotheses were that respondents would be deterred from obtaining definitions if it was hard to access them or if the definitions did not provide useful information, information relevant to the respondent's judgment.

Respondents were able to obtain definitions in one of three ways that varied the number of clicks required: (1) they could display the definition by simply clicking on the highlighted term in the question (one click); (2) they could display the definition by first clicking on the highlighted term, which displayed a list of all terms for which definitions were available, and then by selecting (clicking on) the term of interest in the list (two clicks); or (3) they could display a definition by clicking at least twice, first on the highlighted term which displayed a text file glossary, and then by scrolling the glossary (clicking the scroll bar at least once) to locate the definition of interest. We created "useful" definitions by including some surprising information that might alter respondents' judgments, e.g., the fact that vegetables include French fries. The idea was that respondents might answer differently when they read this kind of definition than when they did not. In contrast, the definitions we created to be "not useful" presented information that was unlikely to affect respondents' answers. Consider the definition for hydrogenated fat: "A fat that has been chemically altered by the addition of hydrogen. Vegetable shortening and margarine are hydrogenated fats." The information in the definition is accurate but not very helpful in evaluating one's consumption of hydrogenated fat.

It is probably not a surprise that respondents tended to ignore the definitions when they had to do something to make them visible. Only 17.4% of the respondents (a total of 501 of them) obtained definitions. This is quite low considering that definitions may, potentially, be essential for respondents to interpret questions in the intended way. The respondents may have been unaware that question terms could have special meanings or the instructions may not have indicated the potential value of the definitions. When respondents did obtain at least one definition, they did so overwhelmingly (89% of the time) for technical terms (e.g. "antioxidants"), where meaning was an obvious concern. Their relatively infrequent requests (11% of the time) for definitions of non-technical terms (e.g. "dairy products") suggest it is easy for respondents to overlook possible differences between their interpretation and the intended one (for example, the definition of dairy products included "cheesy foods like pizza" though many respondents probably would not ordinarily include these). The difference due to the type of terms ($p < .001$) suggests that at least sometimes respondents did not get any definitions because they did not realize they might need them.

Another factor in the low percentage of respondents who accessed definitions was that the amount of effort required; even one click was more than what respondents were willing to expend. Those respondents who did obtain definitions did so far more often when one click was required (56% of the time) than when two or three were required (24% and 20% of the time respectively). The difference due to effort (number of clicks) required ($p < .001$) may indicate that those respondents who never obtained definitions at all were unwilling to invest even one click. The general implication is that interactive features in Web surveys should be designed so that they are very easy to use, requiring no more than one click. When the process involves multiple steps, respondents may begin to invoke the feature, but they are relatively unlikely to complete it. Respondents using the two-click interface started the process by clicking on the highlighted term in the question 629 times but completed it by selecting the term from the list only 246 times. Unless an item is easily seen or it's on the critical path for completing the task, it is unlikely to have much impact on respondents.

Conclusion

Although we've presented quite a few results, we can boil them down to three main themes:

- Respondents attend to pictures and the pictures they see can affect their answers;
- They also attend to the position and spacing of the response options and they use simple heuristics to interpret these and similar visual cues;
- They tend to attend to what's immediately visible and to overlook information they have to make visible.

Consider first the effects of pictures in Web questionnaires. We argue that photographs and other images are powerful contextual stimuli. They can render some instances of a target category more accessible to retrieval than others, leading to assimilation effects. For example, when the pictures display infrequent instances of a category, respondents give lower frequency estimates for the category. By contrast, when the pictures cue more frequent instances, respondents give higher frequency estimates. Similarly, pictures can serve as standards of comparison. Show the respondents a picture of a healthy young woman jogging and they may lower their ratings of their own health. Show them a woman in a hospital bed and it boosts the ratings of their own health. And the effects of pictures are not confined to pictures incorporated into the questionnaire itself, but may extend to photographs intended to create a "human" interface for the survey. The image of a woman professional may subtly alter responses to questions about sex roles.

Our studies also demonstrate the impact of spacing and positional cues on answers. The importance of these and similar cues has been demonstrated repeatedly in the work of Redline and Dillman as well. Like Redline and Dillman (Jenkins & Dillman, 1995; Redline & Dillman, 2002), we find support for the general conclusion that respondents have expectations about the visible aspects of survey questions. They expect a series of items or response options to follow a logical progression from left to right or from top to bottom. They slow down when this expectation is violated. They may infer something about unfamiliar items, such as hotel chains or cars, from their position in a list of similar items. When items are grouped, respondents expect them to be related to each other; as a result, presenting a battery of items on a single screen leads to higher intercorrelations among them then presenting them individually on successive screens (Couper, Traugott, and Lamias, 2001). Finally, respondents expect that the conceptual midpoint of the scale to fall at the visual midpoint. When the visual and conceptual midpoints don't coincide, it throws them off and may affect their answers.

People attend to the information that they see. They give more weight to information that they can see than to information that's not immediately visible. If respondents have to work to see a response option (for example, when a drop box doesn't display all the options initially), they are less likely to select it. If they have to click to see a definition for a key term, they are unlikely to do so; and if they have to click more than once, they are even less likely to bother. It is probably useful to think

of visibility as a continuum, ranging from information that can only be seen with great difficulty to information we can hardly ignore. If we want respondents to attend to something, we need to make it not just visible, but visually prominent. Otherwise, they are likely to ignore that information in favor of information that easily seen.

Acknowledgments: The research described here was supported by an NSF grant (SES 0106222) to Roger Tourangeau, Mick Couper, and Fred Conrad. We gratefully acknowledge NSF's support. (Other contributors for the original NSF award include the Department of Energy, Internal Revenue Service, Social Security Administration, National Agricultural Statistics Service, Bureau of the Census, Department of Transportation, Bureau of Labor Statistics, and Office of Justice Programs. The National Center for Education Statistics contributed funds for a supplemental award.) Much of the work carried out here was conducted in collaboration with staff at MSInteractive. We appreciate the contributions of Reg Baker, Scott Crawford, and Duston Pope at MSIinteractive both to the design and implementation of these studies. We also acknowledge the able assistance of Andrey Peytchev.

References

Burgess, J.H. (1984), *Human factors in forms design*. Chicago: Nelson-Hall.

Couper, M.P. (2001). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, *64*, 464-494.

Couper, M.P., Traugott, M., and Lamias, M. (2001). Web survey design and administration. *Public Opinion Quarterly, 65*, 230-253.

Couper, M.P., Tourangeau, R., & Kenyon, K. (In press). Picture this! Exploring visual effects in Web surveys. *Public Opinion Quarterly*.

Dennis, J.M. (2001). Response timing and coverage of non-Internet households: Data quality in an Internet-enabled panel. Paper resented at the annual meeting of the American Association for Public Opinion Research, Montreal, Canada, May.

Dillman, D.A. (1978), *Mail and Telephone Surveys; The Total Design Method*. New York: Wiley.

Dillman, D.A., Redline, C.D., and Carley-Baxter, L.R. (1999). Influence of type of question on skip pattern compliance in self-administered questionnaires. *Proceedings of the American Statistical Association, Survey Research Methods Section*.

Hoffman, D. (2000), *Visual intelligence: How we create what we see*. New York: W. W. Norton and Company.

Jenkins, C.R. and Dillman, D.A. (1995). Towards a theory of self-administered questionnaire design. In L. Lyberg , P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality*. New York: Wiley.

Kane, E.W., & Macauley, L.J. (1993). Interviewer gender and gender attitudes. *Public Opinion Quarterly*, *57*, 1-28.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in

surveys. *Applied Cognitive Psychology*, *5*, 213-236.

Krosnick, J., and Alwin, D. (1987). An evaluation of a cognitive theory of response order effects in survey measurement. *Public Opinion Quarterly*, *52*, 526-538.

Mangione, T.W. (1995). *Mail surveys: Improving the quality*. Thousand Oaks, CA: Sage.

Redline, C.D., and Dillman, D.A. (2002), The influence of alternative visual designs on respondents' performance with branching instructions in self-administered questionnaires. In R. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (Eds.), *Survey nonresponse* (pp. 179-193). New York: John Wiley

Sanchez, M.E. (1992). Effect of questionnaire design on the quality of survey data. *Public Opinion Quarterly*, *56*, 206-217.

Schwarz, N., and Bless, H. (1992). Scandals and public trust in politicians: Assimilation and contrast effects. *Personality and Social Psychology Bulletin*, *18*, 574-579.

Smith, T.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. 1046-1051.

Sproull, L., Subramani, M., Kiesler, S., Walker, J.H., and Waters, K. (1996). When the interface is a face. *Human-Computer Interaction*, *11*, 97-124.

Tourangeau, R., Couper, M.P., and Steiger, D.M. (2003). Humanizing self-administered surveys: Experiments on social presence in Web and IVR surveys. *Computers in Human Behavior*, *19*, 1-24.

Tourangeau, R., Rips, L.J., and Rasinski, K. (2000). *The Psychology of survey response*. Cambridge University Press.

Walker, J., Sproull, L., & Subramani, R. (1994). Using a human face in an interface. *Proceedings of the conference on human factors in computers* (pp. 85-99). Boston: ACM Press.

Waller, R. (1984). Designing government forms: A case study. *Information Design Journal*, *4*, 36-57.

Wright, P. and Barnard, P. (1975). "Just Fill in this Form" — A review for designers. *Applied Ergonomics*, *6*, 213-220.

Discussion of
The Impact of the Visible

Cleo Redline
National Science Foundation

I am very honored to be here today to discuss the work presented by Tourangeau and his colleagues. Before I talk about the specifics of their research, however, I would like to begin with some background.

Over the years Dillman and I have come to suggest that respondents make mistakes responding to visually-administered questionnaires (that is both paper and Web) not only because they do not understand the verbal language of the questionnaire, but because they do not understand the numeric, symbolic, and graphic language as well. (Redline and Dillman, 2002). The verbal language refers to the words, the numeric, the numbers, the symbolic, the symbols, like arrows, and the graphic language is the conduit by which all of the other languages are conveyed and includes the brightness, color, shape, and location of the information. The reason I have reiterated this framework here is because I used it to process the results of the Tourangeau paper, which I think we can all agree are very impressive and very exciting.

I propose rearranging the original content of the Tourangeau et al. paper. I propose discussing the topic Spacing and Position first and Images as Context last.

Spacing and Position or "Location, Location, Location"

Spacing and position is mostly about manipulating the visual element of 'location,' so conceptually it is simpler or more elemental than the remaining topics. Then within Spacing and Position, I propose leading with positional inferences because it tested the heuristic 'left and top' means first, so conceptually it is a good place to begin. Then the remaining two categories, which demonstrate that the physical middle means midpoint, logically follow.

Left and Top Means First

In 1945 Brandt published an eye-movement analysis using a card with squares that were symmetrically located about a locus (Brandt, 1945). The results of his study may be one of the first to demonstrate the heuristic that the left and top of a space is first because he found that all things being equal, subjects' eyes were naturally attracted to the upper left-hand quadrant, and that the least preferred space was the lower right hand quadrant.

However, as we know, items are NOT of equal interest in a questionnaire. There are what we might call conceptually related zones. The work presented today suggests that respondents attribute meaning to the physical space of a conceptual zone. If one overlays physical quadrants on conceptual spaces, one finds that, unless influenced otherwise, respondents tend to process

information within a conceptual space in a somewhat predictable order, starting at the top, and left and working across then down, and as shown by Tourangeau and his colleagues, they attribute meaning to the order in which they process this information and its position within this space.

However, surveys begin to get complicated, the moment we deviate from using standard text in a standard size and font because it is clear we really don't understand the effect of the other languages yet—a case in point is the divider line that was used to separate the substantive from the nonsubstantive answer categories. The divider line is an example of using the additional language of symbolic language from the Redline and Dillman framework. Tourangeau and his colleagues discovered that not only did the divider line (this symbolic language) influence respondents' answer choices within the substantive range as predicted; it also had the unintended consequence of attracting respondent's attention to the nonsubstantive response options.

We witnessed similar effects in the Census questionnaire. For instance, the population count question, the most critical question on the questionnaire was supposedly ideally positioned in the upper left-hand corner of the questionnaire, under the heading of Step 1, but the results of an experiment informed by cognitive interviews revealed that respondents were drawn to the large write in space for their name instead, which falls much further down the page under Step 2, Person 1 (Dillman et al., 1996). This is an example in which all of the languages (that is, the verbal, symbolic, numeric, and graphic) conspired to draw respondents' attention away from the most critical question on the questionnaire to a question that was of little importance located further down the page. Thus, the 'left and top means first heuristic' can be overruled in ways that we are beginning to identify and slowly come to understand.

Visibility Principle or "Out of Sight, Out of Mind"

The visibility principle is conceptually related to the last topic in that we are still talking about manipulating the visual element of location, it is just that now we are moving things, relocating them, response options in this case, and definitions in the second, so that it requires additional effort on respondents' part to find them.

I learned as a result of eye-movement research with branching instructions that as little as 9 to 12 characters can place information out of view (Redline and Lankford, 2001), so it is not surprising that if information is placed behind closed doors, so to speak, in drop down boxes, or behind links, respondents will be less likely to see it, but it is great to have more experimental confirmation to this effect.

A recent example I have of the visibility principle comes from my work with the Graduate Student Survey at the National Science Foundation. This survey has both a paper and a Web version—and of the 17 interviews I've conducted with this survey, not one person has ever accessed any of the Web definitions, which are hidden behind a help menu that is itself hidden from view in the upper right hand corner.

And the paper questionnaire is barely any better because the definitions are provided in what can only be called a thicket of text—my point here is that things can be hidden when they appear to be in clear view too.

Images as Context or "A Picture is Worth a Thousand Words"

What we learn in this section of the paper is that pictures can prime memories, in a sense they can provide definitions for respondents, or they can act as standards of comparisons. However, when the picture is in the header, assimilation rather than contrast seems to be the result. This made me think about Feynman's description of science as a chess game, the fact that is the scientist's job to figure out the rules of the game from observing it being played. And just when you think you have the rules figured out, someone will castle, and you'll go, 'what was that??!!" I began to wonder if a possible explanation was that the picture in the header was less visible and therefore having less of an effect. It was then that I realized that we needed a no picture condition to compare to, just to rule out this possibility. However, I don't really think that is what this is because I would expect the results to converge in the 'within header' condition. I really think this may be a castle--something exciting to look forward to solving.

Impact on Federal Statistics

Before I conclude, I would like to reflect on how I think the research presented today will impact Federal Statistics. On the one hand, I'm excited because I think it is going to go a long way towards bringing attention to these much-deserved issues. And I also think that, as a result, we will be better poised to conduct high quality household Web surveys, especially attitudinal and behavioral surveys. So I most certainly think we should continue in this vein. However, I must also admit to being concerned, concerned that too much emphasis is going towards the design of Web surveys when paper surveys are still the real work horses.

Take for instance the survey I mentioned earlier, the Graduate Student Survey. It is an establishment survey with both a paper and Web component. The Web component gets an 80 percent response rate, so it was taken for granted that respondents were answering the Web survey—but when I went into the field I discovered that 70 percent of the respondents were actually filling out the paper questionnaire first or performing hand calculations on paper, then simply using the Web as a dissemination tool. Thus, paper remains an important tool for respondents, and the true interface between the survey and the respondent is often the paper questionnaire, not the Web.

Also, it is often said that because of the differences between paper and the electronic medium, paper questionnaires cannot simply be transferred to the Web--that the translation process is simply not a one-to-one mapping (e.g., Murphy et al., 2001). But this philosophy ignores the fact that many of these paper questionnaires are so poorly designed to begin with that one would not want to copy them to the Web as they are. A word of caution I have, therefore, is not to overlook the design of paper surveys in our frenzy to design good Web surveys. It may be that the two will need to work in unison, and if we have people working in isolation on one version or the other, which is the direction I see things moving right now, I think we are headed for trouble.

Conclusion

That word of caution aside, I would like to end by saying that without a doubt this is an exceptional set of experiments that provides a great deal of evidence in support of the notion that the visible matters. Although the experiments were carried out in Web surveys, I have tried to demonstrate that there is every reason to believe that the underlying principles hold true for paper questionnaires too, which I would sum up as: location, location, location; out of sight, out of mind; and a picture is worth a thousand words.

References

Brandt, H. (1945), *The Psychology of Seeing*, New York: The Philosophical Library.

Dillman, D., Jenkins, C., Martin, E., and DeMaio, T. 1996. "Cognitive and Motivational Properties of Three Proposed Decennial Census Forms." Report Prepared for the Bureau of the Census. Washington, D.C.

Murphy, E., Nichols, E., Anderson, A., Harley, M., and Pressley, K. 2001. "Building Usability into Electronic Data-Collection Forms for Economic Censuses and Surveys," *Statistical Policy Working Paper 34—Part 4 of 5*. Washington D.C.: The Federal Committee on Statistical Methodology.

Redline, C. and Dillman, D. 2002. "The Influence of Alternative Visual Designs on Respondents' Performance with Branching Instructions in Self-Administered Questionnaires," in Groves, R., Dillman, D., Eltinge, E., and Little, R. (eds.) *Survey Nonresponse*. New York: John Wiley and Sons, Inc.

Redline, C. and Lankford, C. 2001. "Eye-movement Analysis: a New Tool for Evaluating the Design of Visually Administered Instruments (paper and Web)," *Proceedings of the Section on Survey Research Methods*, American Statistical Association.