Spacing, Position, and Order:
Interpretive Heuristics for Visual Features of Survey Questions

Roger Tourangeau
Mick P. Couper
Frederick Conrad

Institute for Social Research, University of Michigan
Joint Program in Survey Methodology, University of Maryland

September 18, 2003

ACKNOWLEDGEMENT

The work reported here was supported by two National Science Foundation grants — SES-9910882 to Roger Tourangeau and Mick Couper and SES-0106222 to Roger Tourangeau, Mick Couper, Fred Conrad, and Reg Baker. We are grateful to Darby Miller Steiger for her contributions to Experiments 1 and 2 and to Reg Baker, Scott Crawford, Duston Pope, and Andy Peytchev for their contributions to Experiments 3 through 6. Scott Fricker helped with some of the analyses; we gratefully acknowledge his assistance.

Address correspondence to:

Roger Tourangeau
Joint Program in Survey Methodology
1218 LeFrak Hall
University of Maryland
College Park, Maryland 20742

Abstract

We present the results of six experiments that demonstrate the impact of visual features of survey questions on the responses they elicit, the response process they initiate, or both. All six experiments were embedded in Web surveys. Experiments 1 and 2 investigate the effects of the placement of nonsubstantive response options (e.g., "No opinion" and "Don't know" answer options) in relation to the substantive options. The results suggest that when these options are not differentiated visually (by a line or a space) from the substantive options, respondents may be misled about the midpoint of the scale; respondents seemed to use the *visual* rather than the *conceptual* midpoint of the scale as a reference point for responding. Experiment 3, which varied the spacing of the substantive options, showed a similar result. Responses were pushed in the direction of visual midpoint when it fell to one side of the conceptual midpoint of the response scale. Experiment 4 examined the effects of varying whether the response options, which were arrayed vertically, followed a logical progression from top to bottom. Respondents answered more quickly when the options followed a logical order. Experiment 5 examined the effects of the placement of an unfamiliar item among a series of similar items. For example, one set of items asked respondents to say whether several makes and models of cars were expensive or not. The answers for the unfamiliar items depended on the items that were nearby on the list. Our last experiment varied whether a battery of related items was administered on a single screen, across two screens, or with each item on its own screen. The intercorrelations among the items were highest when they were all on the same screen. Respondents seem to apply interpretive heuristics in assigning meaning to visual cues in questionnaires. They see the visual midpoint of a scale as representing the typical or middle response; they expect options to be arrayed in a progression beginning with the leftmost or topmost item; and they expect items that are physically close to be related to each other conceptually.

Self-administered questionnaires have always relied mainly on text to convey the questions and to provide any instructions for answering them. Still, as Redline and Dillman (Dillman, Redline, and Carley-Baxter, 1999; Jenkins and Dillman, 1995; Redline and Dillman, 2002; see also Dillman and Christian, 2003) have pointed out, this verbal language is supplemented by various visual "languages," including symbols (such as arrows and boxes), graphical features (font size, boldfacing), and pictures. The impact of these visual languages on survey responses may be especially marked in Web surveys, because Web surveys often incorporate richer visual material than self-administered paper questionnaires and because the visual presentation of the questions in Web surveys may vary markedly across the range of browsers, screen settings, and user preferences available to respondents.

Unfortunately, the implications of visual presentation of the questions are not especially well understood, even for older methods of data collection. Although several texts offer practical guidelines for the design of mail and other self-administered questionnaires (e.g., Dillman, 1978; Mangione, 1995), apart from the work of Redline and Dillman, there has been relatively little empirical work or theoretical analysis of the issues involved. Most of the work that has been done has focused on designing forms to minimize unit and item nonresponse. For example, the recent work by Redline and Dillman (2002) on the use of arrows and boxes in mail surveys examines navigation problems with contingent questions. But the design of paper forms and computer screens may affect not only whether respondents answer the right questions but also which answers they give (e.g., Sanchez, 1992; Smith, 1995). Smith (1995) cites several examples that demonstrate that unintended variations in printing paper questionnaires can produce dramatic shifts in the answers. Still, the study of forms design is in its infancy, and the impact of forms design on measurement error has been almost entirely neglected. Thus, the rapid spread of the Web as a vehicle for survey data collection has brought to the fore important but overlooked 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 most 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.

This paper focuses on some of the measurement issues raised by Web data collection. It presents the results of six experiments conducted as part of several Web surveys. The experiments are based on the assumption that respondents follow simple heuristics in interpreting the visual features of questions. These interpretive heuristics may sometimes lead to unintended inferences about the meaning of a question, based on incidental visual cues. In the same way, respondents often make inappropriate inferences about the meaning of survey questions based on incidental verbal or numerical cues, such as the numbers assigned to the scale points (Clark & Schober, 1992; Schwarz, 1994, 1996; Schwarz, Grayson, & Knäuper, 1998). Hoffman (2000) argues that interpretive rules are central in visual processing generally and are responsible for such key abilities as depth perception and the perception of shape. The rules for interpreting ordinary visual stimuli can sometimes lead to systematic misinterpretations of those stimuli, producing optical illusions. Similarly, the application of interpretive heuristics for visual cues in questionnaires can lead to misreadings of survey questions.

We distinguish five heuristics that respondents may follow in interpreting visual questionnaires:

- Middle means typical;
- Left and top mean first;
- Near means related;
- Up means good; and
- Like means close.

Each heuristic assigns a meaning to a spatial or visual cue. For example, according to the first heuristic, respondents will see the middle item in an array (or the middle option in a set of response options) as the most typical. Because the middle option is seen as typical, respondents may seize on it as an anchor or reference point for judging their own position (cf. Tourangeau, Rips, & Rasinski, 2000, pp. 244-246). A study by Schwarz and Hippler (1987) provides evidence that respondents see the middle option as

representing the typical or mean value for the population. They asked respondents how much TV they watched in a typical night; respondents reported they watched more TV when the response categories emphasized the high end (the first response category in this condition was "Up to 2 ½ hours") than when they emphasized the low end (e.g., the initial category was "Up to ½ hour"). They also asked respondents about the average value in the population depending on which set of response categories they received and found that respondents' estimates of the population average were affected by the scale range. Respondents seem to make similar inferences about typicality when the options are presented aurally (Tourangeau and Smith, 1996); in Web surveys, visual cues determine which option is seen as representing the midpoint of the scale and, therefore, the typical response.

The second heuristic implies the leftmost or top item in a list of items will be seen as the "first" in some conceptual sense. This interpretive principle reflects the reading order of English (and most Western languages). When the list is a series of ordered response categories or scale values, respondents will expect the topmost or leftmost option to represent one of the two endpoints ("Agree strongly"); in addition, they will expect each of the successive options to follow in some logical order ("Agree", "Neither agree nor disagree") and they will expect the final option in the list — the one in the bottom or rightmost position — to represent the opposite endpoint ("Disagree strongly"). If the list does not conform to these expectations, respondents may become confused, make mistakes, and take longer to respond. As a result, respondents will answer more quickly and more accurately when the order of the response options is consistent with the heuristic than when it is inconsistent with it (Dillman & Christian, 2003). Moreover, when the items in a set do not follow any particular order, respondents may nonetheless infer that they follow a logical progression.

The third interpretive heuristic states that respondents expect items that are physically near each other on the screen (or on the page) to be more closely related conceptually. Respondents will, for example, see two items adjacent to each other on a single screen to be more closely related than the same two items presented on separate screens. They will infer greater conceptual distance between response options that are physically further apart. This heuristic recalls the Gestalt Law of Proximity (Koffka,

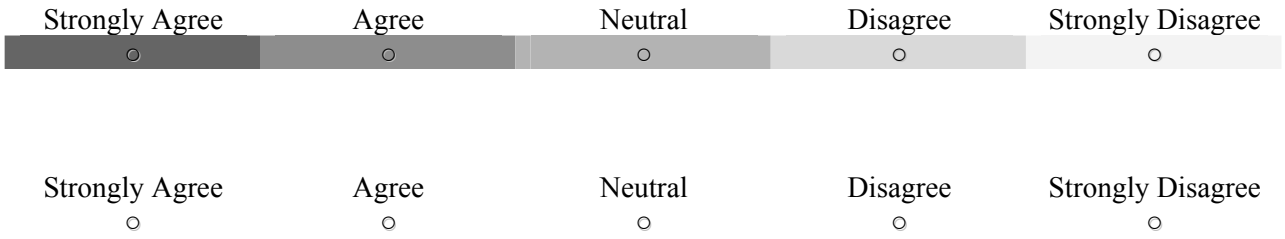
1935; Wertheimer, 1923; see also Kubovy, Holcombe, & Wagemans, 1998), which states that perceivers tend to group nearby objects into figures. We note that the spacing may be a particularly important issue with Web surveys since spacing may accidentally vary because of differences across Web browsers; in addition, the need for respondents to scroll may make spacing more salient in Web surveys than in paper questionnaires. An implication of this heuristic is that items placed on the same screen or in a grid are likely to be seen as more closely related than when they are administered on separate screens; as a result, the answers are likely to be more highly correlated (Couper, Traugott, & Lamias, 2001, provide evidence confirming this prediction).

The fourth interpretive heuristic is a variant on the first; it states that, with a vertically oriented list, the top item or option will be seen as the most desirable. Respondents are most likely to apply this heuristic when the items or options in the list clearly vary in their desirability. Thus, when the list consists of a series of evaluative scale points ("Strongly approve," "approve," and so on), respondents will expect the positive points to come at the top. Similarly, if they are asked their preferences about a series of items displayed in a vertical array, they may expect the best ones to be at the top. Of course, the verbal labels attached to the options or the respondents' actual preferences among the items will generally override the impact of the positional cues, but things will go most smoothly when the positive end of the rating scale (or the most desirable of the items) appears at the top — respondents will answer more quickly and more reliably when the options are positioned in the order suggested by the heuristic. The relation between vertical position and differences in desirability is apparent in large number of deeply-rooted metaphors (e.g., happiness is up, sadness is down; heaven is up, hell is down; good news is upbeat, bad news is downbeat; "on the rise" connotes success, "on the decline" failure; and so on; see Carbonell, 1983).

The final heuristic asserts that visually similar options will be seen as closer conceptually. This heuristic was anticipated by the Gestalt Law of Similarity, which states that perceivers tend to see similar objects as forming a single figure (Wertheimer, 1923). Consider, for example, scales that vary the appearance of the response options (see, for example, Figure 1). According to the "Like means close" heuristic, the endpoints of scales that vary in brightness (like the top one depicted in Figure 1) will be

seen as further apart than the endpoints of a scale that do not vary in brightness (like the bottom scale shown in the figure). This example is analogous to one investigated by Schwarz and his colleagues (Schwarz, Knauper, Hippler, Noelle-Neuman, & Clark, 1991). That study asked respondents to rate their success in life on response scales in which the eleven points were labeled with numbers ranging from -5 to +5 or from 0 to 11. For both groups, the answers tended to pile up at the high end of the scale — most people see themselves as successful — but this tendency was more pronounced when the low end was labeled with negative numbers. Schwarz and his colleagues argue that the negative numbers suggested that the low of the scale was reserved for abject failure whereas the low positive numbers anchoring the low end of the second scale suggested that these points indicated mere lack of success. Both that study and the top example in Figure 1 follow the general principle that when things differ along two dimensions (in the case of the top scale in Figure 1, in both brightness and value), they are seen as differing more sharply than when they differ only in a single attribute (as with the bottom scale). Judd and Harackiewecz (1980) dub this principle the "accentuation effect." When the end points in a scale are perceptually more distinct, respondents will see them as differing more from each other conceptually — that is, as representing more extreme positions. (In the study done by Schwarz et al., the ends of the scale seemed further apart conceptually to the respondents when the numerical labels differed both in sign and value than when they differed only in value.)

Figure 1. Response Scale that Varies Shading of Response Options



This paper presents the results of six experiments embedded in four Web surveys. Each of the experiments tests predictions based on one of the first three heuristics for interpreting visual cues in survey questionnaires. Each experiment examines how the visual appearance or arrangement of the questions affects the response process, the answers to the questions, or both.

SPACING AND THE "MIDDLE MEANS TYPICAL" HEURISTIC

We conducted three experiments to test the implications of the "middle means typical" heuristic. All three experiments tested items in which the response options formed a scale with a conceptual midpoint (that is, a middle option) but that option fell to one side of the *visual* midpoint of the scale. Our prediction is that placing the conceptual midpoint to one side of the visual midpoint will increase the proportion of respondents selecting a response from the *opposite* side, shifting the overall mean of the responses. Put another way, we predicted that responses would be displaced toward the visual midpoint. The three experiments used three different methods to create discrepancies between the visual and conceptual midpoints of the scale.

Experiments 1 and 2: Separating Scale Points from Nonsubstantive Responses

Our first two experiments compared several methods for including nonsubstantive answer categories ("Don't know" and "No opinion" responses) along with substantive scale responses. The first of these experiments compared two versions of a pair of attitude questions. One version presented the nonsubstantive options simply as additional radio buttons. The other version used a divider line to separate the scale points from the nonsubstantive options (see the top panel of Figure 2 below). The second experiment examined the same two attitude items but added one further condition — it examined using spacing to segregate the scale points from the nonsubstantive options. The questions asked whether the federal government was doing enough to ensure equal job opportunities for women and whether it was doing enough to provide day care for working parents. The response scale for the items had a clear conceptual midpoint — "About the right amount." We thought that when the nonsubstantive options

were not visually distinct from the scale points (that is, when they just appeared to be additional radio buttons) respondents would use the visual midpoint of the scale ("Too little") as a reference point and would be more likely to select answers from that conceptual end of the scale.

Method. Both experiments were embedded in Web surveys conducted on our behalf by the Gallup Organization. Survey Sampling Inc. (SSI) selected the samples for both surveys, using a frame (the eLite frame) that consisted of more than seven million e-mail addresses of Web users. The e-mail addresses were provided by visitors to various Web sites who agreed to receive messages on a topic of interest. SSI selected 25,000 e-mail addresses for our first survey and 30,000 for our second and sent out e-mail messages inviting the recipients to take part in "a study of attitudes and lifestyles." The e-mail invitations included the URL of the Gallup Web site with our questionnaire and a unique ID number (which prevented respondents from completing the survey more than once). Gallup's proprietary software administered the Web questionnaires.

The first survey ran from March 28 to April 10, 2001; the second, from February 26, 2002 to March 12, 2002. Of the 25,000 invited to participate in the first survey, 2,987 completed the entire survey (and 255 others got part way through) for a response rate (AAPOR RR1) of 11.9 percent. Of the 30,000 invited to take part in the second survey, 1,590 responded (and an additional 192 began the survey but did not finish it) for a response rate of 5.3 percent.[1] These response rates are quite low by the standards of high-quality mail surveys but are quite typical for Web surveys of this type (see, for example, Couper, 2001).

The questionnaires for the two surveys were quite similar and included questions on a range of topics, including gender attitudes, socially desirable (voting, church attendance) and socially undesirable behaviors (drinking and illicit drug use), impression management, and demographic characteristics. Tourangeau, Couper, and Steiger (2003) provide additional details about the first survey. The key item

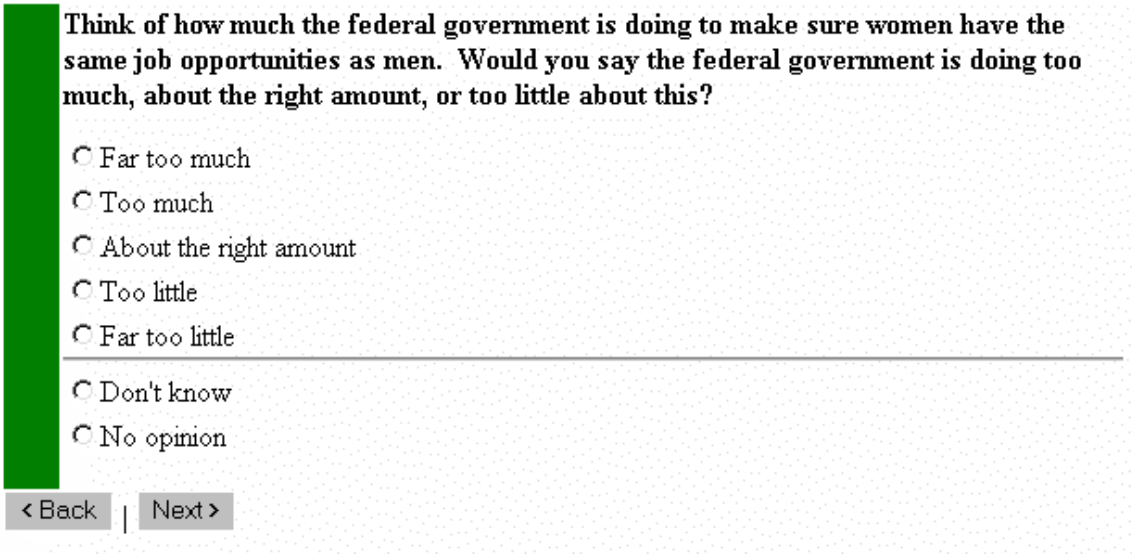
[1] All of the response rates presented here follow the formula for AAPOR Response Rate 1. This formula treats partial respondents as nonrespondents; it also assumes all of the sample cases were eligible for the survey.

shown in Figure 2 came as the tenth question in both surveys.[2] Respondents in the first survey were randomly assigned to receive one of the two versions of the question displayed in Figure 2; those in the second were randomly assigned to the no divider version, a divider line version with a shorter line than the one used in Experiment 1 (one that didn't extend across the whole screen), or a third version in which the nonsubstantive options were separated from the remaining radio buttons by a space. Item 11 in both surveys repeated the experiment with a second question ("Think of how much the federal government is doing to provide daycare centers for the children of working parents. Would you say the federal government is doing too much, about the right amount, or too little about this?"). Respondents received the same version of both questions.

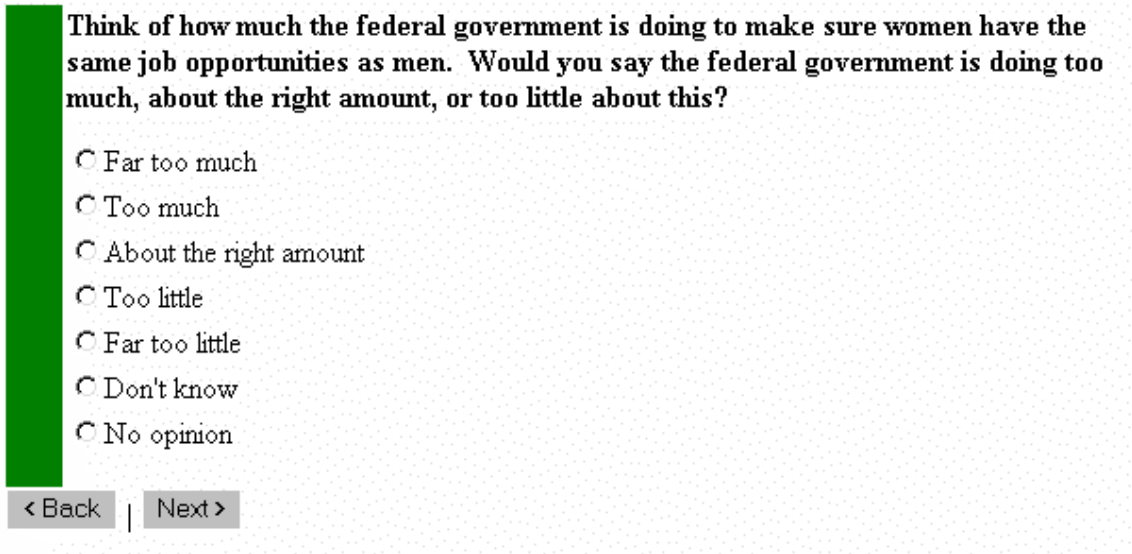
[2] Although occasionally we refer to the question numbers of particular items, none of the surveys reported here actually displayed the question numbers to the respondents.

Figure 2. Formats for Displaying Nonsubstative Options in Experiment 1

a. Divider Line Version



b. Version with No Divider Line



Results. In the versions of the questions that used a divider line or space to separate the nonsubstantive options from the substantive responses, the visual midpoint of the scale fell at the conceptual midpoint ("About the right amount"). In the versions that presented the nonsubstantive options simply as additional radio buttons, however, the visual midpoint falls on one end of the scale ("Too little"). In both experiments, this affected the distribution of the answers, although the results were somewhat more dramatic in the first experiment. Table 1 presents the main results from both items and both studies. (In the second experiment, the results for the divider line and spacing conditions are quite similar and we combine them here.) In all four cases, the means for the substantive answers are higher when there is no separation between the five scale points and the two nonsubstantive answers than when a divider line or spacing separates the scale points from the nonsubstantive answers. (Higher scores represent views more toward the "Far too little" end of the scale; lower scores represent views more toward the "Far too much" end of the scale.) The difference is statistically significant for both items in Experiment 1 and for Question 11 in Experiment 2. There is one additional trend apparent in the data — respondents were more likely to select one of the nonsubstantive options when they were clearly separated from the remaining response options. This pattern is only significant for Question 10 in the first experiment. Apparently, the divider line drew attention to the nonsubstantive options and made respondents more likely to select them.

Table 1. Mean Substantive Response and Percent Choosing Nonsubstantive Answers, by Experiment and Condition

	Experiment 1		**Experiment 2**	
	Q10	Q11	Q10	Q11
Mean Substantive Response				
No Divider	3.69 (1283)	4.01 (1350)	3.51 (431)	3.88 (450)
Divider/Spacing	3.57 (1117)	3.88 (1207)	3.47 (843)	3.75 (885)
	$F(1,2398)=9.20$**	$F(1,2398)=12.3$***	$F(1,1272)<1$	$F(1,1333)=4.49$*
% Choosing DK, No Option				
No Divider	17.5% (1555)	13.0% (1552)	21.1% (546)	16.4% (538)
Divider/Spacing	21.0% (1421)	15.1% (1421)	24.6% (1118)	19.4% (1098)
	$\chi^2=7.24$**, df=1	$\chi^2=2.58$, df=1	$\chi^2=2.56$, df=1	$\chi^2=2.23$, df=1

Note: Parenthetical entries are cells sizes. *** indicates p<.001; ** indicates p<.01; * indicates p<.05.

Experiment 3: Uneven Spacing between Scale Points

We carried out another experiment testing the hypothesis that respondents see the visual midpoint of the scale as representing the conceptual midpoint; again, we predicted that this heuristic would affect their interpretation of the response categories — and the answer they selected. More specifically, Experiment 3 examined what happens when the answer categories for an item are unevenly spaced and, as a result, the conceptual midpoint for an item does not coincide with the visual midpoint. Figure 3 displays an example of even and uneven spacing. When the spacing is uneven, four of the options are to the left of the visual midpoint of the scale and the options on the right appear closer to the midpoint than when the options are spaced evenly.

Figure 3. Evenly and Unevenly Spaced Response Scales

a) Uneven Spacing

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?

Certain	Very likely	Probable	Even chance	Possible	Unlikely	Impossible
○	○	○	○	○	○	○

b) Even Spacing

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?

Certain	Very likely	Probable	Even chance	Possible	Unlikely	Impossible
○	○	○	○	○	○	○

Method. Experiment 3 was embedded in a Web survey conducted by MSInteractive from March 26, 2003 through April 7, 2003. The sample consisted of 39,217 e-mail addresses drawn from two SSI frames for Web surveys. The first frame was the eLite frame used in Experiments 1 and 2; the second was SSI's Survey Spot sample, an opt-in Web panel of almost one million persons who have signed up on line to receive survey invitations. SSI invited sample members to take part via an e-mail invitation that asked them to complete a survey of attitudes and lifestyles funded by the National Science Foundation; the invitation included the URL for the questionnaire. Nonrespondents received one follow-up e-mail. The questions were administered via SPSS's mrInterview software.

The experiment involved two items about the respondents' overall health. The first one asked, "How would you rate your health?" The other item, displayed in Figure 3, asked respondents how likely it was they would get sick enough during the next year that they would 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?"). These were the first two items in the questionnaire. Responses to the item asking for ratings of overall health were unaffected by the spacing of the response alternatives; we focus on responses to the second item.[3]

Results. We expected that the respondents who got the unevenly spaced options (shown in the top panel of Figure 3) would be more likely to select answers from the right side of the scale (the unlikely end) than those who got the evenly spaced options (shown in the bottom panel). In the uneven spacing condition "Possible" and "Unlikely" are closer to the visual midpoint and are, therefore, more likely to be selected than in the even spacing condition. The results are in line with that expectation. With the uneven spacing, 63.4 percent of the respondents chose answers from the right side of the scale ("Possible," "Unlikely," or "Impossible"); with the even spacing, the percentage dropped to 58.3 (χ^2 =

[3]The overall health question followed one of two pictures — a photograph of a healthy young woman jogging or 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. Neither the photograph nor its position affected responses to the item on chances for a day in bed due to illness.

7.54, df = 1, $p < .01$). An analysis of variance indicates that the overall means of the responses were also affected; on the average, responses indicated less chance of a sick day in bed when the response options were unevenly spaced (mean of 4.60) than when the were evenly spaced (4.45) — $F(1,2689) = 6.78$, $p < .01$.

ORDER AND THE "LEFT AND TOP MEAN FIRST" HEURISTIC

Another heuristic respondents may use in understanding response options and selecting an answer from among them is the "Left and top mean first" heuristic. According to the heuristic, the leftmost or top item in a list of items should be the "first" in some conceptual sense and the remaining options should follow from left to right or from top to bottom in some logical progression. For example, scale options often follow a graded sequence starting at one pole, proceeding through the intermediate values, and ending at the other pole. When the scale options do not, in fact, follow their conceptual rank order, it will slow respondents down (and may affect their answers). Moreover, when a series of items seems to follow some order, respondents may use the position of an unfamiliar item within the series to infer its characteristics. We carried out experiments to test both these predictions.

Experiment 4: Order of the Response Options

We varied the order of the response options in four behavioral frequency items and two attitude items. The frequency items used a graded frequency scale ranging from "Never" to "Every day." Similarly, the attitude items used a five-point agree-disagree scale. One version of the items presented the response options in an order that was consistent with the "Top is first" heuristic; that is, the top option was one of the endpoints ("Never," "Strongly agree") and the each of the succeeding options followed in order of extremity. A second version of the items departed somewhat from this order, putting "Never" and "It depends" at the bottom of the scale. The final visual departed more sharply from the order prescribed by the heuristic. This version of the frequency items presented "Never" as the topmost option, followed by "Every day," with the remaining options in order of decreasing frequency. The final version

of the attitude items presented "It depends" as the top option, followed by "Strongly agree," "Strongly disagree," "Agree, and finally "Disagree." Table 2 displays the three versions of the response scales for the agree-disagree items. Our main hypothesis is that greater departures from the order implied by the heuristic would lead to slower response times.

Table 2. Response Scales in Experiment 4

Consistent with Heuristic	Mildly Inconsistent	Strongly Inconsistent
○ Strongly agree	○ It depends	○ It depends
○ Agree	○ Strongly agree	○ Strongly agree
○ It depends	○ Agree	○ Strongly disagree
○ Disagree	○ Disagree	○ Agree
○ Strongly disagree	○ Strongly disagree	○ Disagree

Method. Experiment 4 was part of a Web survey conducted by MSInteractive from April 2 through April 23, 2002. Like the sample for Experiment 3, the sample for this survey consisted of e-mail addresses drawn from the eLite and Survey Spot SSI frames. SSI sent invitations to 14,192 e-mail addresses drawn from the two frames. 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 completing it, for a response rate of 18.1% (not counting the partials). Again, the mrInterview software administered the questionnaire. Among other experiments, the survey included several comparing response options in three different orders.

We carried out two independent experiments, one with four frequency items and the other with two agree-disagree items. The four frequency items asked how often the respondent took vitamins (Question 1), ate fruit (Question 2) and pasta (Question 3), and went shopping (Question 15). The agree-disagree questions (Questions 13 and 14) asked about following the doctor's advice ("It is SENSIBLE to do exactly what the doctors say") and about when the respondent went to the doctor ("I have to be VERY ILL before I go to the doctor"). We compared three versions of both sets of questions (see Table 2). As

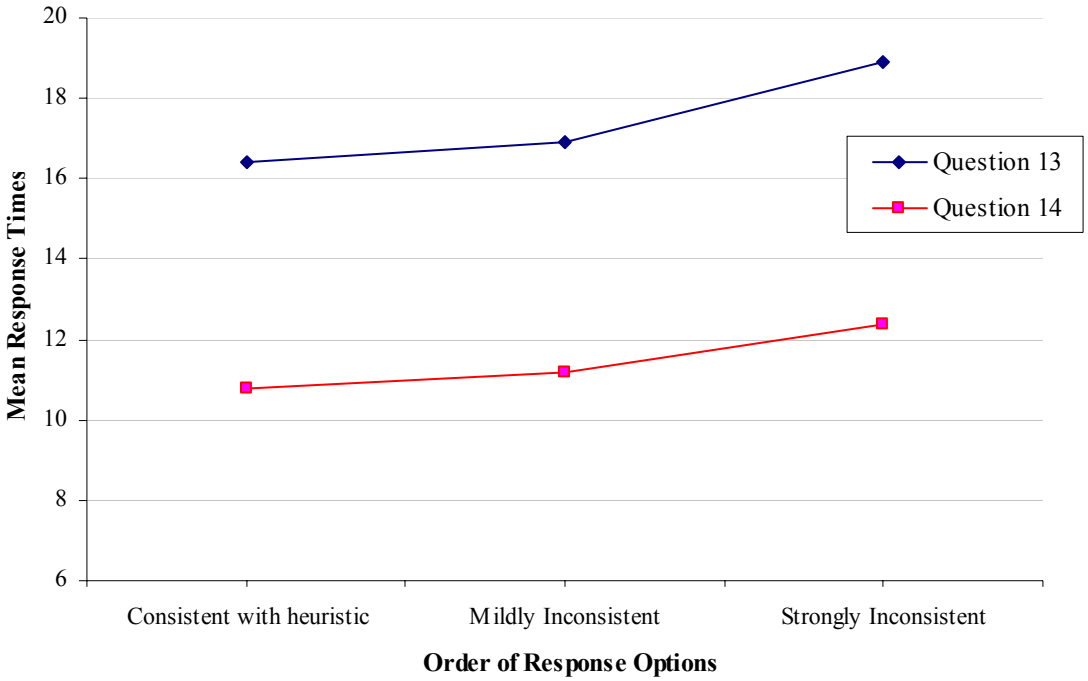
noted earlier, one version of the items presented the response options in the order implied by the "Top means first" heuristic; the other versions departed mildly or sharply from that order. All three versions arrayed the options vertically.

Respondents got the response options in the same order for all four of the frequency items; similarly, they got the same version of both agree-disagree items. (There were independent random assignments of the respondents to the versions of the two sets of items.) The program recorded both answers to the items and the time that elapsed from the time of presentation of the item to the response.[4]

Results. We anticipated that respondents would answer the questions most quickly when the items followed the order implied by the "Top means 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). After dropping unusually long response times (those over 60 seconds), we compared mean response times across the three versions of the items. Three of the six items showed significant differences in response times, but the results were clearest for the two agree-disagree items. Figure 4 below displays the average response times for those two items. For both, the differences in response times across experimental treatments were highly significant: $F(2,2533) = 18.7$ for Question 13 (on following the doctor's advice) and $F(2,2591)=12.6$ for Question 14 (on going to the doctor). The degrees of freedom for the two items vary because of item nonresponse. For Question 13, response times increased from an average of 16.4 seconds for the version in which the options followed the order consistent with the heuristic to a mean of 18.9 seconds in the conditions in which the options were most inconsistent with the heuristic. For Question 14, the mean response times went from 10.8 to 12.4 seconds across the three versions. The response times for the four frequency items show no consistent pattern.

[4]Experiments 4-6 were carried out as part of two Web surveys that each included a large number of independent experiments. A danger with such designs is that the experimental manipulations, which were designed to affect one or two nearby items, may have had remote effects on later items or interacted with variables manipulated later in the questionnaire. Because we carried out randomization separately for each of the experiments, we were able to look for such carryover effects. In 154 tests focusing on the outcomes in Experiments 4, 5, and 6, we found 10 significant effects; this is a rate of 6.5 percent, not much greater than the 5 percent false positive rate to be expected by chance.

Figure 4. Response Times and Consistency with Heuristic



The rearrangement of the response options also affected the distribution of responses. Again, the results were most striking for the two agree-disagree questions. The proportion of respondents selecting the "It depends" option dropped dramatically when that option came at the bottom of the list (in the second version of the items) than when it came in the middle or at the top of the list (in the first and third versions). With Question 13, the proportion choosing the "It depends" option fell from 45.8 percent (when it was the middle option) or 41.5 percent (when it was the top option) to 20.8 percent (when it came last). These differences across version were highly significant — $\chi^2 = 127.4$, df = 2, $p < .001$. The comparable figures for Question 14 were 27.7 percent (when "It depends" was the middle option), 21.8 percent (when it was the top option), and 8.2 percent (when it came at the bottom of the list). Again, the differences were highly significant ($\chi^2 = 106.0$, df = 2, $p < .001$). For both items, then, the "It depends" option is most popular when its position suggests that it represents the midpoint of the scale.

Discussion. The findings for the two agree-disagree items supported our predictions, but those for the four frequency items were mixed at best. One difference between the two sets of items may account

for the difference in the results. Frequency questions generally presuppose that the respondent engages in the behavior in question — takes vitamins, eats pasta, or goes shopping. Because the "Never" option denies this presupposition, it may be seen as a special option — that is, as not part of the scale. Thus, whether it comes at the top or the bottom and whether it is placed next to the frequent or infrequent options may have relatively little impact on how respondents process it.

Experiment 5: Positional Inferences

When confronted with an unfamiliar item, respondents may use the "Top means first" heuristic to infer the item's characteristics from its position within a series of similar items. For example, consider respondents who have been asked to judge the prices of a series of cars. If the cars seem to follow some order (say, from expensive to inexpensive models), then the respondents may use this order in judging where an unfamiliar make and model (say, the Fiat Tipo) falls on the spectrum. We carried out an experiment involving six sets of items that tested whether respondents drew such inferences based on positional cues.

Method. This experiment was embedded in the same Web survey as Experiment 3. It included six target items, each of them embedded in a series of seven related items. Within each of the series, we varied the position of one of the items (a relatively unfamiliar one) so that it appeared either as the third or seventh in the list; the other items were arranged to convey the sense that they were presented in rank order from top to bottom. In each case, we hoped that the six nontarget items would form a Guttman scale with the nontarget items arrayed in order by scale position. For each set of items, the majority of the respondents did, in fact, give answers on the six nontarget items that conformed to the requirements of a Guttman scale. For example, respondents judged whether each of seven makes and models of car (BMW 318, Acura Integra, Fiat Tipo, Mazda Protégé, Toyota Corolla, Dodge Neon, and Geo Metro) were expensive. The highest proportion of respondents rated the BMW as expensive, the next highest proportion the Acura, and so on down the list. Fiat Tipo was either the third car from the top or at the

bottom of the list. Each series was presented in a grid format, with each item presented in one row of the grid.

The six target items asked whether isoflavin was important for a healthy diet (one of the seven dietary items in Question 7), whether cod liver oil was low in saturated fat (one of seven fats in Question 8), whether Clarion Inns were expensive hotels (one of seven hotel chains in Question 18), whether the cost of living was high in Ocala, Florida (one of seven cities in Question 19), whether the Austin Rover was an expensive midsize car (one of seven midsize cars in Question 20), and whether the Fiat Tipo was an expensive small car (one of seven small cars in Question 21). All of the items were yes-no items.

Results. Table 3 displays the main results from this study, the proportion of respondents answering "yes" for each of the target items. The vertical position of the item made a significant difference for five of the six items. The differences are particularly large for judgments of the Clarion Inns and the Fiat Tipo, both of which were more likely to be seen as expensive when they came near the top of the list (among the more expensive hotels and cars) than when they came at the bottom (among the cheaper ones).

Table 3. Proportion Yes (and Sample Size), by Position in Grid

Judgment/Item	**Percent Yes** (n)		
	Third from Top	Bottom	χ_1^2
Important for healthy diet/Isoflavin	46.8 (1325)	45.7 (1253)	<1
Low in saturated fat/Cod liver oil	43.8 (1361)	37.8 (1287)	9.69**
Expensive hotel/Clarion Inn	63.3 (1357)	45.8 (1285)	82.0***
Expensive city/Ocala, FL	53.0 (1357)	60.4 (1298)	14.9***
Expensive midsize/Austin Rover	93.8 (1368)	88.7 (1285)	22.2***
Expensive small car/Fiat Tipo	74.9 (1352)	62.8 (1275)	44.4***

Note: *** indicates p<.001; ** indicates p<.01; * indicates p<.05.*

We classified each judgment according to whether it was consistent with the intended judgments for the items presented at the top of each series. For example, we classified the judgment that the Clarion Inns were expensive as consistent with the top hotels in the array of hotels. (And, in fact, the items at the top of each list did elicit very high levels of consensus, with 97 percent or more of the respondents giving the same judgment for the top item in each series). Across the six judgments, the respondents who got the target items in the third position in the series were significantly more likely to give a judgment consistent with the item at the top of the list than those who got the target items in the bottom position — 62.8 percent versus 58.9 percent; $F(1,2720)=21.1$, $p<.001$. Still, the results were in the expected direction only for four of the six target items, and the two reversals are statistically significant. They involve the judgments of cod liver oil and Ocala, Florida. Cod liver oil was *more* likely to be rated low in saturated fat when in came at the beginning of the series (just below lard and butter) than when it came at the bottom (just below olive oil and light vegetable spread). Similarly, Ocala, Florida, was *less* likely to be seen as an expensive city to live in when it came near the top of the list (just below San Francisco and Houston) than when it came at the bottom (after Bismarck, North Dakota, and Springfield, Illinois).[5]

Discussion. Although the overall trend for the six target items is consistent with the "Top means first" heuristic, there do seem to be systematic differences across items. The heuristic assumes that, when respondents are unfamiliar with a specific item but the series seems to be arrayed in order along the dimension of interest, they will use the position of the item within the series to infer its value. It is possible that respondents knew enough about the two items for which reversals were found that they contrasted them with nearby items. For example, respondents may have inferred that Ocala must be at least somewhat expensive to live in simply because it is in Florida; when Ocala came just below the most expensive cities in the series, it apparently seemed less expensive to the respondents than when it came just below cities that are relatively inexpensive places to live (like Bismarck and Springfield). Similarly,

[5] We also looked at the impact of the position of the target item on ratings of the other six items in each set. Of the 36 nontarget items across the six sets, only three showed significant differences depending on the position of the target item within the array. These differences didn't follow any clear pattern.

respondents may have inferred that cod liver oil must be relatively low in saturated fat (after all, it is taken from a fish) and contrasted it with lard and butter when it came just below those items on the list.

Still, even in the cases for which we observed reversals, position within the list clearly mattered. Whether respondents assimilate the target item to nearby items (as we expected) or contrasted it with its neighbors, judgments of the target item were clearly affected by the local context and this context was determined by the physical position of the target item within the series.

EXPERIMENT 6: GROUPING AND THE "NEAR MEANS RELATED" HEURISTIC

Still another heuristic that respondents may apply in interpreting Web questions is the "Near means related" heuristic. Based on this heuristic, respondents expect items that are physically near each other on the screen to be closely related conceptually. We tested one implication of the heuristic — that respondents will see stronger interconnections among items that are displayed on a single screen than among those displayed on separate screens, boosting the correlations among them.

Method. This experiment was embedded in the same Web survey as Experiment 4. It compared the presentation of a battery of eight related items on a single screen versus presentation across several screens. These items were presented about midway through the questionnaire as the twelfth question. All eight questions used the same seven-point response scale (with the endpoints labeled "Agree" and "Disagree") and all of them asked about diet:

A. Maintaining a healthy diet is a priority in my life.
B. I avoid "fast food" because it's not healthy.
C. I monitor my cholesterol level closely.
D. How food tastes is more important to me than its nutritional value.
E. I pay attention to nutritional information on food packaging.
F. I limit the amount of red meat in my diet.
G. My lifestyle makes it difficult to eat right.
H. I try to balance my diet across the key food groups.

The first version presented the eight items on a single screen in a grid format. The second version presented the items on two screens, both as four-item grids. (One grid presented items A—D above; the

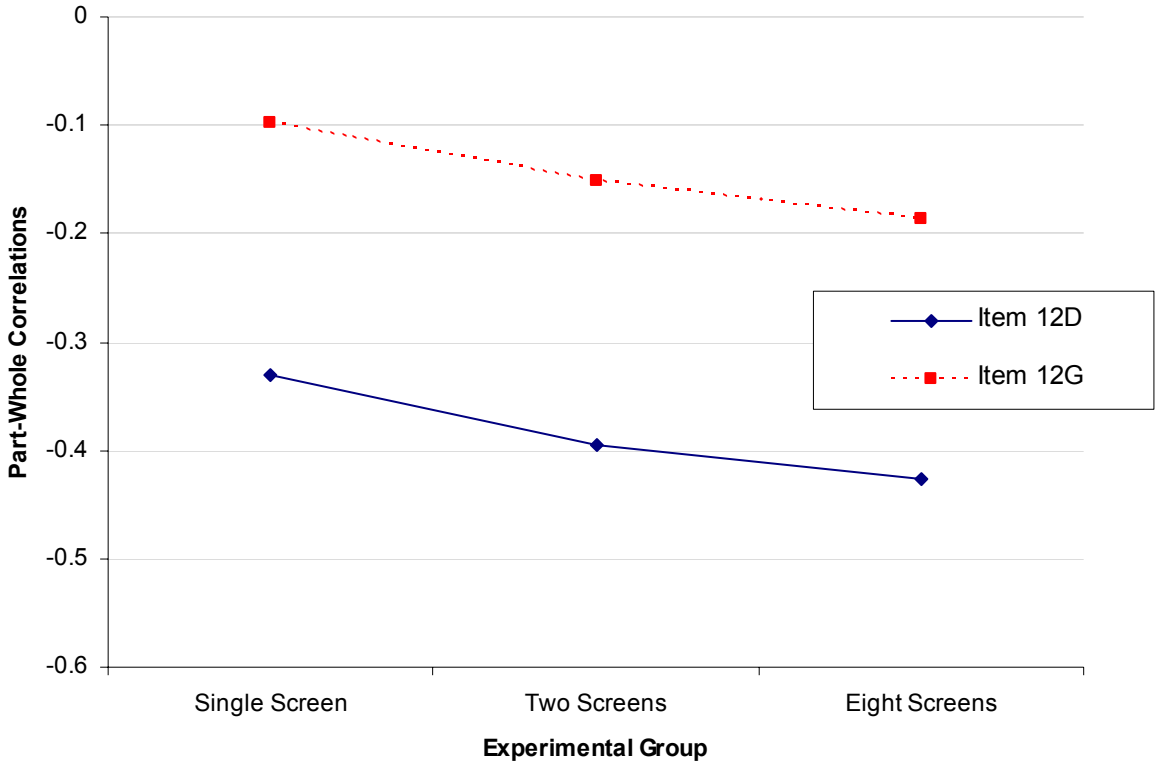
other items E—F). The final version presented one question per screen. We tested the hypothesis that the single grid design would encourage greater consistency among the answers. This hypothesis reflects the "Near means related" heuristic; thinking that the items are more closely related when they are presented on the same screen, the respondents will tend to answer them more consistently.

Results. As we expected, the responses to the eight diet questions were more highly intercorrelated when the items were presented in a grid on a single screen (Cronbach's alpha of .621) than when the eight items were presented in two grids on separate screens (alpha of .562) or when each item appeared on its own screen (Cronbach's alpha of .511). If we treat the alpha values as correlations and apply Fisher's z transformation to them, then the linear trend across the three conditions is significant ($z = 2.62$, $p < .01$). In addition, when the eight items appeared in two grids of four items each, the median correlation for items in the same grid was .382, but the median correlation was only .351 for items in different grids. These findings are consistent with the "Near means related" heuristic and they replicate the findings of Couper, Traugott, and Lamias (2001) on grids versus single items.

Higher correlations do not necessarily mean more valid responses. We looked at the extent to which respondents gave the same answers to all eight items, a response tendency that Krosnick (1991) has termed "nondifferentiation." More specifically, we calculated the proportion of items for which each respondent gave his or her most common response (cf. Holbrooke, Green, and Krosnick, 2003). With eight questions and seven response categories, this proportion could range from .25 to 1.0 for respondents who answered all eight questions. (Respondents who answered all eight questions would have selected the same answer category at least twice, producing a minimum score of .25.) Respondents who got the items on a single screen showed less differentiation than those who got them on two screens or on eight screens (mean nondifferentiation scores of .436, .422, and .412, respectively; $F(2,2519)=6.11$, $p<.01$). Two of the items (items D and G above) were "reverse worded." For these two questions, agreement indicated *less* concern about diet; for the other six, agreement indicated greater concern. The relation of these two items to overall scale scores (the part-whole correlations) was lowest when the eight items were presented in a single grid ($r = -.331$ for Question 12D and -.097 for Question 12G) than were they were

presented in two grids on separate screens (r's of -.395 and -.151) or on eight separate screens (-.427 and -.187). Apparently, respondents were less likely to notice the reverse wording when the items appeared in a single grid. These results are graphed in Figure 5.

Figure 5. Part-Whole Correlations for Reversed Worded Items, by Format of the Eight-Item Battery



We also compared the time it took for respondents to answer all eight questions. We set very long response times (those greater than 240 seconds overall for the eight items) to missing and then carried out a one-way analysis of variance on the remaining times.[6] Respondents took less time on average to answer the eight questions when they were presented in a grid on one screen (a mean of 60.4 seconds) than when they were presented in two grids on different screens (65.4 seconds) or individually on eight separate screens (99.0 seconds) — $F(2,2493) = 322.1$, $p < .001$.

[6] We dropped a total of 71 response times, or about 2.8 percent of the times. The results do not change appreciably if we delete only those times greater than 480 seconds.

Discussion. When the items were grouped in a single grid, respondents seemed to infer greater similarity among them than when they were spread across two or eight screens; in fact, they may have inferred more similarity among them than was warranted. Respondents were more prone to select the same answer for all eight items and they seemed less sensitive to the fact that two of the items were reverse worded when the eight items were in a single grid. Respondents seemed to use the proximity of the items as a cue to their meaning, perhaps at the expense of reading each item carefully. There was also a large difference in response times across the three experimental groups — almost 40 seconds; doubtless, much of this difference reflects the time it took to download the items. It is also possible, though, that the single-grid format encourages respondents to move quickly through the items quickly, making it more likely they will overlook some of the distinctions among them.

GENERAL DISCUSSION

We find evidence in all six of our experiments that respondents make inferences about the meaning of survey items based on visual cues such as the spacing of the response options, their order, or the grouping of questions. These inferences affect how quickly respondents answer the questions, which answers they select, or both. In Experiments 1 and 2, respondents seemed to use the visual midpoint of the scale as a reference point for choosing their answers. When the nonsubstantive response options (such as "Don't know") were integrated visually with the scale points (as in the bottom panel of Figure 2), shifting the visual midpoint, the answers shifted as well. Respondents were more likely to select a response from the bottom end of the scale — that is, their answers shifted in the same direction as the visual midpoint. Similarly, when we altered the spacing of the answer categories in Experiment 3 (see the top panel of Figure 3), the answers on the right side of the scale (which were now closer to the visual midpoint) became more popular than they were with equal spaced response categories. We argue that respondents apply a heuristic in which the visual midpoint is seen a providing a benchmark, representing either the conceptual midpoint of the scale or the most typical response.

Respondents also seem to expect items to be arrayed in a logical progression from left to right or from top to bottom. In a survey that uses aural administration of the questions, respondents may have similar expectations regarding successive items, but in a Web survey (or a self-administered paper questionnaire) the cues that trigger these expectations are visual rather than temporal. Our experiments provided two kinds of evidence for the "Top and left mean first" heuristic. We presented respondents with questions in which the answer categories systematically violated the heuristic (Experiment 4). This slowed them down and affected the distribution of their answers; in line with the heuristic, they were more likely to pick the neutral "It depends" option when it came in the middle of the scale. In addition, Experiment 5 showed that respondents inferred something about the characteristics of an unfamiliar item by its position in an array of similar items. Across six parallel items, the position of the unfamiliar item affected the ratings it received; apparently, the placement of the item in the array conveyed information about the position of the item on the dimension of judgment. For a couple of the items, the effect seemed to work in the opposite direction, with respondents contrasting an item with nearby items on the list. In part, these reversals may have occurred because the item seemed out of order — that is, because respondents inferred where it should have been placed in the array.

Another cue that affected respondents' answers was the grouping of the questions onto Web pages. Experiment 6 compared responses to a battery of related questions on diet when the items were presented in a single grid on one screen, in two grids on separate screens, or one item at a time across eight screens. We expected that respondents would see the items as more closely related when they all were placed together in a grid and that is what we found — the average intercorrelation of the items (as measured by alpha) went up when the eight items were placed together in the same grid. The inference that the items were closely related because they were visually grouped may have supplanted a more careful reading of the items. Respondents were more prone to use the same answer from one item to the next and were less sensitive the fact that two of the items were worded in the opposite direction from the other six when all eight items were placed together in a grid.

These studies extend two earlier lines of research. First, they provide additional evidence that the visual layout of survey items matters. Work by Smith (1994), Sanchez (1992), and Redline and Dillman (2002) have shown similar effects for paper questionnaires to the ones we demonstrate here for Web surveys. Our studies are based on somewhat different premises from the earlier studies; our studies assume that respondents make quick interpretive judgments about the questions based on the visual cues in the questionnaire, much as the interpretation of purely visual stimuli is based on the rapid application of interpretive rules (Hoffman, 2000). Our studies also extend earlier work on the (unintended) effects of verbal cues on how respondents interpret and answer survey questions (Clark & Schober, 1992; Schwarz, 1992, 1996; Schwarz et al., 1991). Respondents are sensitive to the numerical labels assigned to the response scale, to the specific values mentioned in the response options, and to other verbal cues. Our results indicate that they may also make unintended inferences based on the visual cues offered by the question. Basing their reading on the question's visual appearance, respondents may miss key verbal distinctions and interpret the questions in ways the survey designers never intended.

REFERENCES

Carbonell, J. (1983). "Derivational Analogy in Problem Solving and Knowledge Acquisition." In R.S. Michalski (ed.), *Proceedings of the International Machine Learning Workshop* (pp. 12-18). Urbana, IL: Department of Computer Science, University of Illinois at Urbana-Champaign.

Clark, H.H., & Schober, M.F. (1992). "Asking Questions and Influencing Answers." In J. M. Tanur (ed.), *Questions about Questions: Inquiries into the Cognitive Bases of Surveys* (pp. 15-48). New York: Russell Sage.

Couper, M.P. (2001), "Web Surveys: A Review of Issues and Approaches." *Public Opinion Quarterly*, 64 (4), 464-494.

Couper, M.P., Traugott, M., and Lamias, M. (2001),"Web Survey Design and Administration." *Public Opinion Quarterly,* 65 (2): 230-253.

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

Dillman, D.A., and Christian, L. (2003), "The Effects of Graphics, Symbols, Numbers, and Words on Answers to Self-Administered Questionnaires." Paper presented at the Annual Conference of the American Association for Public Opinion Research, May 18, Nashville, Tennessee.

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* (pp. 743-748).

Holbrook, A.L., Green, M.C., and Krosnick, J.A. (2003), "Telephone vs. Face-to-face Interviewing of National Probability Samples with Long Questionnaires: Comparisons of Respondent Satisficing and Social Desirability Bias." *Public Opinion Quarterly,* 67 (1): 79-125.

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 et al. (eds.), *Survey Measurement and Process Quality*. New York: Wiley.

Judd, C., and Harackiewicz, J. (1980), "Contrast Effects in Attitude Judgment: An Examination of The Accentuation Hypothesis," *Journal of Personality and Social Psychology*, 38: 390-398.

Koffka, K. (1935). "Perception: An Introduction to the Gestalt-theorie," *Psychological Bulletin*, 19: 531-585.

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

Kubovy, M., Holcombe, A.O., and Wagemans, J. (1998). "On the Lawfulness of Grouping by Proximity." *Cognitive Psychology*, 35 (1): 71-98.

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 (2): 206-217.

Schwarz, N. (1994), "Judgment in a Social Context: Biases, Shortcomings, and the Logic of Conversation." *Advances in Experimental Social Psychology*, 26: 123-162.

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

Schwarz, N., Grayson, C.E. and Knäuper, B. (1998), "Formal Features of Rating Scales and the Interpretation of Question Meaning." *International Journal of Public Opinion Research*, 10 (2): 177-183.

Schwarz, N., and Hippler, H.-J. (1987), "What Response Scales May Tell Your Respondents: Information Functions of Response Alternatives." In H.-J. Hippler, N. Schwarz, and S. Sudman (eds), *Social Information Processing and Survey Methodology*. New York: Springer-Verlag.

Schwarz, N., Knäuper, B., Hippler, H.-J., Noelle-Neumann, E. and Clark, F. (1991), "Rating Scales: Numeric Values May Change the Meaning of Scale Labels." *Public Opinion Quarterly*, 55 (4): 618-630.

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.

Tourangeau, R., Couper, M.P., & 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.

Tourangeau, R., and Smith, T.W. (1996). "Asking Sensitive Questions: The Impact of Data Collection Mode, Question Format, and Question Context." *Public Opinion Quarterly*, 60 (2): 275-304.

Wertheimer, M. (1923). "Untersuchen zur Lehre von der Gestalt II." *Psycologische Forschung*, 4: 301-350. (Translation published in 1938 as "The Laws of Organization in Perceptual Forms" in W. Ellis [Ed.]), *A Sourcebook of Gestalt Psychology* [pp. 71-88]. London: Routledge & Kegan Paul.)