“UP MEANS GOOD”

THE EFFECT OF SCREEN POSITION ON EVALUATIVE RATINGS IN WEB SURVEYS

ROGER TOURANGEAU*

MICK P. COUPER

FREDERICK G. CONRAD

Abstract This paper presents results from six experiments that examine the effect of the position of an item on the screen on the evaluative ratings it receives. The experiments are based on the idea that respondents expect “good” things—those they view positively—to be higher up on the screen than “bad” things. The experiments use items on different topics (Congress and HMOs, a variety of foods, and six physician specialties) and different methods for varying their vertical position on the screen. A meta-analysis of all six experiments demonstrates a small but reliable effect of the item’s screen position on mean ratings of the item; the ratings are significantly more positive when the item appears in a higher position on the screen than when it appears farther down. These results are consistent with the hypothesis that respondents follow the “Up means good” heuristic, using the vertical position of the item as a cue in evaluating it. Respondents seem to rely on heuristics both in interpreting response scales and in forming judgments.

ROGER TOURANGEAU is a Vice President at Westat, Rockville, MD, USA, and co-director of the Survey Methodology Unit there. MICK P. COUPER and FREDERICK G. CONRAD are Research Professors at the University of Michigan’s Survey Research Center, Ann Arbor, MI, USA, and at the Joint Program in Survey Methodology at the University of Maryland, College Park, MD, USA. The work reported here was supported by grants from the National Science Foundation [SES-0106222 to R.T., M.P.C., F.G.C., and Reg Baker] and the National Institute for Child Health and Human Development [R01 HD041386-01A1 to R.T., M.P.C., F.G.C., and Reg Baker]. Neither the National Science Foundation nor the National Institute for Child Health and Human Development is responsible for the conclusions presented here. We are grateful to Reg Baker for his contributions to the design and implementation of these experiments, and to Jenna Fulton, Hanyu Sun, and Cong Ye for their work on the meta-analysis. *Address correspondence to Roger Tourangeau, Westat, 1600 Research Boulevard, Rockville, MD 20850, USA; e-mail: RogerTourangeau@westat.com.

doi:10.1093/poq/nfs063

© The Author 2013. Published by Oxford University Press on behalf of the American Association for Public Opinion Research. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Introduction

Survey respondents sometimes have trouble determining the intended meaning of survey items, and they may draw on apparently incidental features of the questions to help them interpret the question or the answer options. For example, Schwarz and his colleagues have shown that the numerical labels assigned to the response scale can have a big impact on the distribution of the answers (Schwarz, Grayson, and Knäuper 1998, experiment 1; Schwarz et al. 1991; see also O’Muircheartaigh, Gaskell, and Wright 1995 and Tourangeau, Couper, and Conrad 2007). Their work demonstrates that it matters whether the numerical labels on the scale points range from -5 to $+5$ rather than from 1 to 11.

Tourangeau, Couper, and Conrad (2007) found that a visual cue—the shading of the response options—can have effects similar to those of positive and negative numerical labels. Their experiments varied whether the response options for a set of items were shaded with a single hue or with different hues at the two ends of the scale (e.g., shades of blue at one end and shades of yellow at the other). The distribution of the answers shifted when the points at opposite ends of the scale were shades of a different hue rather than shades of a single hue. Like the use of positive and negative numbers, the use of different hues seemed to signal greater distance between the two ends of the scale and reduced the number of respondents selecting answers from the low end. The findings from that study also suggested that there is a hierarchy of interpretive cues, with verbal labels for the scale points taking precedence over numerical labels and numerical labels taking precedence over color (see Toepoel and Dillman [2008] for further evidence on the relative priority of different cues).

Tourangeau, Couper, and Conrad (2004, 2007) argue that respondents apply five heuristics in interpreting questions in Web surveys. The five heuristics assign meanings to different visual features of the questions:

- 1) Middle means typical;
- 2) Left and top mean first;
- 3) Near means related;
- 4) Up means good; and
- 5) Like means close.

Their experiments to date have examined four of the five heuristics. This paper presents evidence for the remaining one, the “Up means good” heuristic.

The first heuristic involves inferring that the visual midpoint of the scale is special in some way, representing either the conceptual midpoint of the underlying dimension (for example, the neutral point of a bipolar scale) or the most typical response. Tourangeau and his colleagues (2004, experiment 3) showed that, when the visual midpoint did not coincide with the conceptual midpoint of the scale (for example, because the scale options weren’t evenly spaced), the distribution of the answers changed. According to the second heuristic (“Left and top mean first”), respondents expect the response options to follow some

logical order (such as progressing from one end of the relevant dimension to the other), beginning with the leftmost option (when the options are arrayed horizontally) and progressing to the rightmost, or beginning with the top option (when the options are arrayed vertically) and progressing to the bottom. When the order of the answer options is inconsistent with the heuristic, response times are much slower than when the order of the answer options is consistent with it (Toepoel and Dillman 2008; Tourangeau, Couper, and Conrad 2004, experiment 4). The third heuristic (“Near means related”) means that respondents may infer a conceptual relation between two items based on their physical proximity. Items in a grid, for example, may be seen as more closely related than the same items presented on different screens (Tourangeau, Couper, and Conrad 2004, experiment 6; for related findings, see Couper, Traugott, and Lamias [2001] and Toepoel and Dillman [2008]). The fifth heuristic (“Like means close”) refers to the tendency for respondents to infer conceptual similarity from similarity in visual appearance. The results on the colors of rating scales illustrate this heuristic. When the two ends of the response scale are shades of the same hue, respondents seem to infer that the two extremes are closer conceptually than when the two ends of the scales are shades of different hues. As a result, the answers may shift (Tourangeau, Couper, and Conrad 2007).

The “Up means good” heuristic involves inferring something about the value of an item from its position on the screen. Why would the position of an item on the screen affect the evaluations it receives? One possibility is that the link between evaluation and vertical position is an example of “grounded cognition.” Many cognitive scientists argue that cognition is often closely linked with perception or with action. For example, according to Barsalou (2008), “when people process animals conceptually, visual areas [of the brain] are especially active; when they process artifacts, motor areas become active” (p. 627). For example, when people see a picture of a hammer, the motor area of the brain involved in grasping a hammer becomes active. Along similar lines, Wells and Petty (1980) argue that there is the same sort of automatic link between the physical act of nodding and the cognitive act of agreeing. Perhaps there is a similar action-cognition association that links looking up at something with having positive feelings for it.

Another possibility is that the connection between evaluation and screen position is purely conceptual, perhaps based on the numerous metaphors that link good things with higher vertical positions (e.g., heaven is above, hell below; upswings are good, downswings bad; an upbeat mood is positive, a downbeat mood, negative; and so on; see Carbonell [1983]). Abstract concepts are often grounded in concrete metaphors (Lakoff and Johnson 1980). Life, for example, is often seen as a journey; the passage of time is often represented as movement through space—and affective experiences are often linked with vertical position. Barsalou (2008) notes that “increasing evidence suggests that these metaphors play central roles in thought.” If the views of grounded cognition theorists like Barsalou are right, the “Up means good” heuristic may

be a manifestation in surveys of a deeply rooted cognitive link between positive affect and vertical position.

Still another possibility is that the more positive evaluations associated with being higher up on the screen reflect the greater ease of processing items that appear near the top of the screen—that is, reflect their increased perceptual fluency (Reber, Winkielman, and Schwarz 1998). Perceptual fluency refers to how easily a stimulus can be seen and identified; several studies show that increased perceptual fluency leads to enhanced liking for a stimulus (Lee and Labroo 2004; Reber, Winkielman, and Schwarz 1998). It may be easier for respondents to see and identify a word or phrase when it appears near the top of the screen than when it appears farther down, and the increased perceptual fluency from being at the top of the screen may lead to more positive ratings.

Whatever the underlying basis for this heuristic, there is some evidence that people use it. In a laboratory study conducted by Meier and Robinson (2004), participants were quicker to classify positive words (*brave*, *loyal*, *hero*) as positive when they appeared near the top of the computer screen than when they appeared near the bottom. The reverse was true for negative words (*bitter*, *clumsy*, *crime*), which participants classified more quickly when they appeared near the bottom of the screen than near the top. Apparently, the participants expected the positive words to be up (and the negative ones to be down), and they were slower to respond when these expectations were not met.

This paper reports evidence for the application of the “Up means good” heuristic in surveys. We describe a total of six experiments designed to test the effect of screen position on evaluative ratings. The studies use different types of samples (some volunteer and some probability samples), different methods to vary the screen position of the items, different items, and differences in screen appearance. Our main purpose in these experiments was to provide evidence for the “Up means good” heuristic and to assess its robustness across these procedural variations.

Method

All six of our experiments were embedded in Web surveys that included other questions (and other methodological experiments). Most of the experiments (experiments 1–3 and experiment 6) use samples recruited from the members of opt-in Web panels. In all but one of these studies, we drew our samples from panels maintained by two different vendors. We used two Web panels in these surveys to assess the robustness of the findings across different panels. Because we found no differences across panels, we present the results for the full samples from these experiments, pooling respondents from the different panels. The other two experiments were conducted with members of a probability Web panel; respondents in these experiments completed similar surveys administered a month apart.

All six experiments varied the position of the target items on the screen, using one of two methods. In experiments 1, 4, 5, and 6, we presented two items on each screen, with one of the items appearing near the top of the screen and the other item farther down; these experiments randomly varied which item appeared in which position. Three of these experiments (4, 5, and 6) also varied how far down the screen the second item appeared—near the middle of the screen or closer to the bottom. In experiments 2 and 3, we presented only one item per screen and varied the position of the item by putting it above or below the response scale. This was useful because it allowed us to manipulate the screen position of an item without pairing it with a second item (thus, avoiding any potential question-order effects).

All of the experiments also varied additional variables that we thought might affect the size of the basic screen-position effect. These included the order of the response categories (positive options first or negative options first) and whether the response options were arrayed vertically or horizontally. We varied the order of the response options in experiments 1 through 5 and the orientation of the options in experiments 1, 4, and 5. We found no systematic interactions between these variables and the position of the target item on the screen, and we ignore these variables below. With one exception, respondents were randomly assigned to an experimental condition by the program administering the Web survey when they got to the relevant section of the questionnaire. The exception occurred in experiments 4 and 5, where the same participants took part in both experiments (about a month apart). They remained in the same condition in experiment 5 that they had been randomly assigned to in experiment 4.

EXPERIMENT 1

Our first experiment was part of a Web survey conducted for us by Market Strategies, Inc. (MSI). The data were collected from June 16, 2006, through July 11, 2006, from members of the Survey Sampling International (SSI) Survey Spot panel, an opt-in Web panel of volunteers who had signed up online to receive survey invitations. The standard SSI incentive (a monthly sweepstakes drawing for cash prizes totaling \$10,000) was offered to those who completed the survey. SSI sent e-mail invitations to 82,585 panel members asking them to take part and, in addition, sent nonrespondents one follow-up e-mail. A total of 2,385 panel members completed the survey. Six hundred twenty-one (621) of those invited to participate, and 240 of those who completed the questionnaire, had taken part in one of our earlier studies (and were, in effect, part of a two-wave panel study). The overall participation rate (see [AAPOR 2011](#), p. 38) was 2.9 percent. Another 546 panel members began the survey without completing it; we discard the data from these break-off cases.¹

1. Including the data from the break-off cases did not alter any of the major conclusions presented below.

In experiment 1, respondents were asked to rate the scientific community, the U.S. Congress, doctors, and managers of health maintenance organizations (HMOs): “The first few items are about various groups that can affect health care. Please indicate how you feel about these groups.” For each item, the response scale consisted of five verbally labeled options (*Very positive*, *Somewhat positive*, *Neither positive nor negative*, *Somewhat negative*, or *Very negative*).

We displayed the four target items on two screens. Approximately half of the respondents got the more popular groups at the top of the two screens (*The scientific community* on the first screen and *Doctors* on the second) and the less popular groups beneath them (*The U.S. Congress* on the first screen and *Managers of HMOs* on the second) as the second item on the screen. The remainder of the respondents got the less popular groups on top. (Our guesses about how our respondents would view these groups were confirmed by the respondents’ ratings.) These four items were the first questions in the survey.

EXPERIMENT 2

We were concerned that the results of experiment 1 might have reflected the strong views respondents had about the various groups they rated in that study. In experiment 2, we asked respondents to rate more neutral objects—specifically, six food ingredients (*riboflavins*, *cornstarch*, *antioxidants*, *flour*, *niacin*, and *all-spice*). These six items came immediately after an introduction: “The next few questions list several food ingredients that might be good or bad for you. For each food ingredient, please rate how good or how bad you think the item is for you.”

This experiment also used a different method for varying the screen position of the items, presenting each target item (e.g., *riboflavins*) either above the response scale (and hence near the top of the screen) or below it. We expected that the vertical position of the items on the screen would affect the ratings they received, especially for the unfamiliar items (such as *riboflavins* and *niacin*). Figure 1 shows screen captures for the main experimental conditions in experiment 2.²

This experiment was part of a second Web survey done for us by MSI. The target items were questions 36 through 41 in the survey, although the item numbers were not displayed to the respondents. The survey was conducted from August 29, 2007, through September 13, 2007. The sample for this survey was drawn from the SSI Survey Spot panel and the e-Rewards panel (an opt-in panel with 3.2 million members). SSI invited 99,381 panelists to take part; 1,200 of them completed the survey (and another 373 gave up part way through without completing it). A total of 31,918 members of the e-Rewards

2. Experiment 2 also varied the response format, comparing a vertical slider bar (shown in figure 1 below) with radio buttons. We thought that, because respondents had to manipulate the slider bar up and down, it might heighten the effect of the item’s position on the screen. (Ten cases did not have Java-enabled browsers, and these were all assigned to the radio button condition.) We found no effects for the format of the response options and pooled the data across the two response formats.

a) Please rate how good or how bad you think the item is for you.

Corn Starch

1 - Very bad for you

2

3

4

5

6

7 - Very good for you

b) Please rate how good or how bad you think the item is for you.

Corn Starch

1 - Very bad for you

2

3

4

5

6

7 - Very good for you

Please rate how good or how bad you think the item is for you.

Corn Starch

Next

c) Please rate how good or how bad you think the item is for you.

Corn Starch

1 - Very bad for you

2

3

4

5

6

7 - Very good for you

Please rate how good or how bad you think the item is for you.

Corn Starch

Next

d) Please rate how good or how bad you think the item is for you.

Corn Starch

1 - Very bad for you

2

3

4

5

6

7 - Very good for you

Please rate how good or how bad you think the item is for you.

Corn Starch

Next

Figure 1. Screen shots for items from experiments 2 and 3. Panels a and b display the slider bar conditions, with the item to be rated placed above (panel a) or below (panel b) the slider bar. Panels c and d show the same conditions with the radio button format. Experiment 3 differed slightly; it did not use numerical labels.

panelists were invited to complete the survey, and 1,206 of them did. (An additional 508 broke off before completing the survey.) Across both samples, the participation rate was 1.8 percent. Again, we discarded the data from cases who broke off without completing the survey.

EXPERIMENT 3

Our third experiment was a replication of the second. Experiment 3 asked about the same six food ingredients as experiment 2 and followed the same experimental design. The items also appeared in the same position (questions 36 through 41) in the two questionnaires. We did alter one feature of the design—we dropped the numbers from the response scales in experiment 2.

Like the second, this experiment was part of a Web survey carried out for us by MSI, this time from April 29, 2008, through May 14, 2008. The sample for this survey was again drawn from two sources, with half of the respondents ($n = 1,200$) coming from the SSI Survey Spot panel and the other half ($n = 1,200$) drawn from the four million members of the Authentic Response Web survey panel. Like the Survey Spot panel, the Authentic Response panel uses a sweepstakes as an incentive. Overall, the participation rate was 5.0 percent; the participation rate for the Authentic Response panel was a bit higher (6.7 percent) than that for the SSI panel (4.0 percent). Again, we discarded the data from the 563 cases who broke off without completing the survey, and we combined the data from the two samples.

EXPERIMENTS 4 AND 5

Experiments 1–3 were all done with members of volunteer Web panels, and we were concerned that our results might apply only to the members of such panels. We therefore carried out two experiments on the members of an Internet panel selected through area probability sampling. The members of this panel—the Face-to-Face Internet Survey Platform (FFISP)—were recruited by interviewers from Abt/SRBI. The sample was a three-stage area probability sample, with metropolitan areas and non-metropolitan counties selected in the first stage of selection, blocks in the second stage, and addresses in the third stage. The addresses came from the U.S. Postal Service’s Delivery Sequence File in urban areas and from enumerator listings in rural areas. One thousand adults were recruited for the panel initially. Our experiments appeared in months 9 and 10 of the panel—that is, in July and August 2009. The overall unweighted response rate for the initial 1,000 recruits (AAPOR RR4) was 42.5 percent ([Sakshaug et al. 2009](#)). Nine hundred thirty-five panel members completed the July questions (for a response rate of 39.7 percent, or 93.5 percent $\times$ 42.5 percent), and 913 completed the August questionnaire (for a response rate of 38.8 percent).

Experiments 4 and 5 used the same food items as experiments 2 and 3 but employed a different method from the earlier studies for varying the

screen position of the items. Respondents got two items on a single screen, and the experiment varied which one appeared at the top of the screen and which one appeared farther down. We paired *Riboflavin* with *Cornstarch*, *Antioxidants* with *Flour*, and *Niacin* with *Allspice*; for each pair, approximately half of the respondents got the first item in the pair at the top of the screen and the rest got the second member of the pair at the top. The assignment of one member of the pair to the top position on the screen was done independently for each pair of items. In addition, the position of the

Figure 2. Screen shots for two of the conditions in experiments 4 and 5. The top panel (panel a) shows the condition in which the second item (Riboflavin) appears about mid-screen; the bottom panel (b) shows the condition in which the second item appears even farther down the screen.

second item was varied experimentally; that item appeared either at about the middle of the screen (see the top panel of [figure 2](#) below) or closer to the bottom of the screen (bottom panel of [figure 2](#)). Respondents got the second item in the same position (near the middle of the screen or near the bottom) for all three pairs (second item near the middle or near the bottom). Respondents received the same items (in the same conditions) in both the July and August surveys. The treatments used in the two studies were identical.

EXPERIMENT 6

We did one final experiment, again collecting ratings of the six food items, but this time adding six physician specialties (pediatrician, gynecologist, internist, endocrinologist, nephrologist, and urologist). The introduction to these items stated: “The next few items are about different medical practitioners. Please indicate how positive or negative you feel about each one.” This experiment varied item position using the same method used in experiments 4 and 5—that is, we varied which item appeared first on the screen and whether the second item appeared near the bottom of the screen or near the middle. The six food items were the first six items in the survey and were followed by the six physician items. Each food item was paired with another food item (e.g., *Riboflavin* and *Cornstarch* were a pair), and one member of the pair was randomly assigned to the top position on the screen. Similarly, each physician specialty was paired with another specialty and one member of the pair was assigned to the top position. Independent randomizations were done for each pair of food items and for each pair of physician specialties. Respondents got the second item in the same screen position (either near the middle of the screen or near the bottom) for all twelve items.

The experiment was part of a Web survey done by MSI, from June 23, 2010, through July 21, 2010. As with experiments 2 and 3, the sample for this study was drawn from two panels, with roughly half coming from the SSI Survey Spot panel ($n = 1,201$) and the other half drawn from the Authentic Response Web survey panel ($n = 1,206$). Overall, the participation rate was 1.6 percent, with a 7.8-percent participation rate for the Authentic Response panel members and a 0.9-percent participation rate for the Survey Spot panel. Again, we discarded the data from the break-offs.

SUMMARY

The six experiments we conducted varied in a number of ways—who collected the data, what panels the respondents came from, what items they rated, and how we varied the screen position of the items. [Table 1](#) below summarizes

Table 1. Selected Characteristics of Experiments about Screen Position, by Experiment

	Sample source (and size)	Items (response categories)	Position manipulation
Experiment 1	SSI Survey Spot (2,385)	Doctors, HMO managers, Congress, the scientific community (Very positive, Somewhat positive, Neither positive nor negative, Somewhat negative, Very negative)	Item first vs. second on the screen
Experiment 2	SSI Survey Spot (1,200) e-Rewards (1,206)	Riboflavin, niacin, antioxidants, flour, allspice, cornstarch (Seven-point scale, with the ends labeled “Very good for you” and “Very bad for you”)	Item to be rated above or below rating scale
Experiment 3	SSI Survey Spot (1,200) Authentic Response (1,200)	Same six food items as in experiment 2 (and same response scale)	Item to be rated above or below rating scale
Experiments 4 and 5	FFISP (935 in July, 913 of them reinterviewed in August)	Same six food items as in experiment 2 (and same response scale)	Item first or second on the screen (second item near middle or near bottom of screen)
Experiment 6	SSI Survey Spot (1,201) Authentic Response (1,206)	Same six food items as in experiment 2 (and same response scale); six physician specialties—pediatrician, gynecologist, internist, endocrinologist, nephrologist, urologist (five-point scale, with the ends labeled “Very positive” and “Very negative”)	Item first or second on the screen (second item near middle or near bottom of screen)

NOTE.—Sample sizes given in terms of respondents who completed the survey. Market Strategies conducted all of the studies, except for experiments 4 and 5, which were done by Abt/SRBI.

the key features of each study. Experiments 1, 4, 5, and 6 presented two items per screen and varied which one appeared near the top; Experiments 4–6 allowed us to compare the ratings each item received in three different screen positions—near the top, near the middle, and near the bottom of the screen. Experiments 2 and 3 presented only one item on each screen and manipulated screen position by varying whether the item to be rated appeared above or below the response categories.

Results

Although we have carried out separate analyses for the results of each experiment, the simplest way for us to present all of the results is to describe the meta-analysis we performed; it examined the results from all six experiments at once and allowed us to see whether any overall pattern emerged. For each item in each experiment, we calculated an effect size. The effect size was the difference between the mean rating an item received (for example, item i from study j) when it was presented in a higher position on the screen ($\bar{X}_{Hij}$) and when it was presented in a lower position ($\bar{X}_{Lij}$), divided by the pooled standard deviation of the ratings (S_{pij}) for that item:

$$ES_{ij} = \frac{\bar{X}_{Hij} - \bar{X}_{Lij}}{S_{pij}}. \quad (1)$$

This is a standard effect-size measure (see, for example, [Lipsey and Wilson 2001](#), pp. 48–49).

A positive effect size means that the item received more favorable ratings on average when it was presented higher up on the screen than when it was presented farther down; a negative number means that the item got less favorable ratings on average when it was presented in the higher position than in the lower one. Standardizing the differences between means (by dividing by the pooled standard deviation for the item) makes it possible to combine the effect sizes across items and across experiments. In experiments 1–3, there were only two screen positions (and therefore only one effect size for each item), but in experiments 4–6, there were three screen positions and thus three possible effect sizes—one comparing the average ratings an item got when it was presented near the top of the screen versus when it was presented near the middle (top versus middle), one comparing the ratings an item got when it was presented near the top of the screen versus when it was presented near the bottom (top versus bottom), and one comparing the ratings an item got when it was presented near the middle of the screen versus when it was presented near the bottom (middle versus bottom). We computed all the effect sizes collapsing across other factors in the study (e.g., whether the response scale was displayed vertically or

horizontally). In computing the mean effect size for each study ($\overline{ES}_j$, the mean of the effects for study j), we used weighted averages:

$$\overline{ES}_j = \frac{\sum_i w_{ij} ES_{ij}}{\sum_i w_{ij}}, \text{ and} \quad (2)$$

$$w_{ij} = 1 / se_{ij}^2.$$

where each individual effect size (ES_{ij} , defined earlier in equation 1, is effect size i from study j) receives a weight (w_{ij}) that was the inverse of its squared standard error (that is, the weight was the error variance for that effect size). This weighting procedure gives more weight to more precise effect-size estimates and less weight to more variable ones. The weight attached to the study-level mean ($\overline{ES}_j$) was the inverse of the average of the individual effect variances (the se_{ij}^2 s in equation 2).

After presenting the basic meta-analytic results, we consider the issue of whether the position of the item on the screen interacted with the other experimental variables or with the background characteristics of the respondents. We then examine the effects of screen position on two additional variables—response times and the consistency of answers over time.

OVERALL RESULTS FOR RATINGS

Table 2 shows the mean effect sizes for each experiment.³ The table also displays the study-level weight assigned to each experiment. The study-level weight is the inverse of the squared standard error of the mean effect size for that study. The grand mean presented at the bottom of the table is a weighted average of all the effects (based on the study-level weights), and it gives more weight to the experiments that yielded more precise means. For example, experiment 1 yielded a mean effect size of 0.160 (that is, about a sixth of a standard deviation); the effect sizes from that study receive relatively high weight in the calculation of the overall mean, because the effect sizes from experiment 1 had relatively small standard errors.

The mean effect sizes that are presented in the left panel of the table are based on all the effect sizes. The means on the right panel of the table are based only on the effect sizes comparing the ratings an item received when it was near the top of the screen with the ratings it got in all other screen positions. This includes all the effect sizes for experiments 1–3, but only two of the three types of effect sizes computed for experiments 4–6 (that is, the means for the final three experiments include the top versus middle and top versus bottom comparisons, but not the middle versus bottom comparisons). We thought the effects might be weaker when the higher position was near the middle of

3. These data are available in an online appendix (please see the [supplementary data online](#)).

Table 2. Mean Effect Sizes and Weights, by Comparison, Model, and Study

	All effects				Top vs. other positions			
	Mean	<i>n</i> of effects	Weight (fixed)	Weight (random)	Mean	<i>n</i> of effects	Weight (fixed)	Weight (random)
Experiment 1	0.160	4	632.7	469.6	0.160	4	632.7	383.3
Experiment 2	0.044	6	297.8	256.0	0.044	6	297.8	228.0
Experiment 3	0.052	6	318.0	270.7	0.052	6	318.0	239.6
Experiment 4	0.055	18	57.3	55.6	0.050	12	95.1	86.6
Experiment 5	0.003	18	56.2	54.5	0.001	12	93.4	85.2
Experiment 6	0.028	36	254.2	223.1	0.030	24	430.6	298.4
	Grand mean	<i>n</i> of effects	Standard error	<i>Z</i>	Grand mean	<i>n</i> of effects	Standard error	<i>z</i>
Fixed	0.088	88	0.025	3.52***	0.079	64	0.024	3.44***
Random	0.083	88	0.027	3.02**	0.074	64	0.028	2.67**

p* < 0.01; *p* < 0.001

the screen rather than near the top, and we dropped these effects from the right panel of [table 2](#).

We also looked at the data using two different models, the fixed effects model and random effects model. The random effects model assumes that in addition to random sampling error, the effect size estimates also vary because of essentially random differences across the studies. The fixed effects model leaves out this study-level variance component. Because the random effects model includes an extra variance term, the study-level variances are larger and the study-level weights smaller under that model. We prefer the random effects model because it better reflects the variation across experiments. Regardless of which model we adopt, there is a significant effect of screen position—the smallest of the four z -values is 2.67, and the corresponding p -values are all less than 0.01.⁴ Items at the top of the screen are rated about 0.08 of a standard deviation higher than those in the middle or toward the bottom. The effect of appearing in the middle versus the bottom of the screen is smaller, but still in the expected direction. The mean effect size for this comparison is 0.03 (data not shown).

We also carried out heterogeneity tests to determine whether the results varied significantly across the six experiments. Looking at all effects (that is, the left panel of [table 2](#)), the test of heterogeneity is not significant ($\chi^2(5) = 5.66$; $p > 0.30$). Similarly, for just those comparisons involving the top of the screen versus other screen positions (the right panel of the table), the heterogeneity test is again not significant ($\chi^2(5) = 6.48$; $p > 0.25$). Thus, the tests indicate that the six experiments give similar results.

[Figure 3](#) shows all sixty-four effect sizes for the comparisons involving the top of the screen versus other screen positions. The bulk of the effect sizes are positive (indicating more favorable ratings when an item appeared near the top than in the other positions on the screen). The mean effect size seems to be largest for experiment 1 and smallest for experiment 5 (the second interview with the FFISP panel members), but, as we noted, despite the many differences between the experiments, the results do not show significant heterogeneity across studies. Several of the effect sizes are negative. We can discern no clear pattern to these cases. Many of the negative effects involve relatively unfamiliar targets, such as *Managers of HMOs* (experiment 1), *antioxidants* (experiments 4 and 5), and *endocrinologists* (experiment 6). Still, some of the negative effects involve the ratings of *flour* (experiments 2 and 3), hardly an unfamiliar item. In addition, some of the largest effects involve *riboflavin* (these are the largest effects in experiments 5 and 6), which also seems a relatively unfamiliar item.

Interactions of screen position with other variables: We carried out a large number of analyses examining the interaction between screen position and

4. We also analyzed the data treating each effect size individually with the studies as clusters; the results of this analysis were quite similar to those presented in [table 2](#). The mean effect size was 0.088, and the t -value was 2.54 (with 5 df , $p = 0.052$)

Figure 3. Effect sizes for each experiment. The effects compare ratings for an item at the top of the screen with ratings for the same item in other screen positions. The x's represent individual effect sizes; the dark squares represent the mean effect sizes for each study.

the background characteristics of the respondents, in particular their age and education. Although we found a few scattered interactions between respondent age and screen position and between respondent education level and screen position, we do not find any consistent pattern to these interactions. We also computed separate effect sizes for respondents who had a college degree and for those with less education. We found no overall difference in the mean effect sizes for the two groups. Younger respondents and older respondents and more educated respondents and less educated ones seem about equally prone to the effects of the position of the items on the screen.

Response times: In experiments 2 and 3, we could compare the time it took to rate the same item when it appeared above or below the response scale. On average, putting the item below the scale increased response times by about a second—10.2 seconds versus 11.3 seconds in experiment 2 ($F(1,2388) = 7.69$, $p < 0.01$) and 9.3 versus 10.4 in experiment 3 ($F(1,2380) = 20.9$, $p < 0.001$).⁵

5. We also examined the possibility that the order of the response options might affect how quickly respondents answered. Experiments 1–5 all varied whether the positive or negative end of the scale came first; experiments 1, 4, and 5 also varied whether the response options were arrayed vertically or horizontally. We found no evidence of an effect of the order or orientation of the response options on response times.

Table 3. Correlations across Waves (Experiments 4 and 5), by Item and Item Position

Item	Item first on the screen	Item second on the screen
Riboflavin	0.607 (443)	0.759 (448)
Cornstarch	0.737 (451)	0.523 (448)
Antioxidants	0.760 (448)	0.775 (431)
Flour	0.774 (430)	0.730 (451)
Niacin	0.656 (437)	0.603 (440)
Allspice	0.571 (443)	0.588 (438)

NOTE.—Parenthetical entries are sample sizes.

Consistency across waves: Almost all of the participants in experiment 5 had also taken part in experiment 4 and, as a result, we can see how the position of the item on the screen affected the consistency of the ratings it received across the two waves of data collection. (Respondents got each item in the same position in both waves.) Table 3 shows the correlations of the answers, by item and screen position. There does not seem to be a clear trend. For three of the items, the across-wave correlations are higher when the item is higher up on the screen, and for the other three, the across-wave correlations are lower when the item is higher up on the screen. The differences by screen position in the across-wave correlations are significant for the first two items in the table (both $p < 0.001$), but not for the other four.

Discussion

Across our six experiments, we found evidence that the position of an item on the screen significantly affected its perceived desirability. Overall, the results indicate that items were rated more favorably (by about 0.08 of a standard deviation) when they appeared near the top of the screen than when they appeared lower down (see table 2); this represents a shift of about 0.10 to 0.15 of a scale point. The effect of screen position seemed a bit stronger in experiment 1 and a bit weaker in experiment 5 (see figure 2), but a test of the heterogeneity of the results finds no overall difference across the six studies. The effect of screen position on evaluative ratings may be small, but it is reasonably robust across differences in samples, differences in the items, and differences in the methods for manipulating screen position. The analyses also show that the effect was not limited to any specific age or education group.

We believe that the impact of the position of an item on the screen on the ratings it receives reflects the use of screen position as a cue. Prior work suggests that survey respondents draw on a variety of cues to help them assign meaning to response scales; they may also use similar cues in making their

judgments. The cues they use include the numbers assigned to the response options (Schwarz, Grayson, and Knäuper 1998, experiment 1; Schwarz et al. 1991; Tourangeau, Couper, and Conrad 2007), the spacing and shading of the options (Tourangeau, Couper, and Conrad 2004, 2007; see also Toepoel and Dillman 2008), and the relative width of the options (Schwarz, Grayson, and Knäuper 1998, experiment 2). In addition, respondents seem to infer the value of the item on the dimension of judgment from its position within a series of similar items (Tourangeau, Couper, and Conrad 2004, experiment 5). In non-survey settings, people infer that the first person to speak is the one most likely to remember an event correctly (Wright and Carlucci 2011) and tend to see the first of two similar paintings or songs as the more attractive (Pandelaere, Millet, and Van der Bergh 2010). The results of our six experiments suggest that respondents use the position of an item on the screen as a cue to help them judge its value. The “Up means good” heuristic may reflect a deeply rooted conceptual link between position and evaluation, a link reflected in many common metaphors (Carbonell 1983). Such links between physical and spatial concepts, on one hand, and subjective concepts, on the other, figure prominently in theories of grounded cognition. Another possibility is that being higher on the screen makes an item easier to see and this enhanced perceptual fluency leads to more favorable ratings (Reber, Winkielman, and Schwarz 1998). Our experiments were not intended to determine the exact basis for the heuristic. However, our finding that respondents took more time to answer (about an additional second) when the item was placed farther down on the screen than when it appeared near the top provides suggestive evidence that is consistent with the perceptual fluency account.

The results from our six experiments also support some additional conclusions. First, respondents pay attention to the visual details in Web and other surveys, sometimes in ways that the researchers may not intend. To paraphrase Smith (1995) slightly, “little things do matter.” Even among the highly experienced Web survey panelists who completed most of our surveys, the position of an item on the screen had systematic (if not always large) effects on the answers. Second, these studies (see also Tourangeau, Couper, and Conrad 2004, 2007) extend earlier findings by Schwarz (1996), who argued that respondents try hard to interpret survey questions (and the associated answer categories) and sometimes base their interpretations on incidental cues. Apparently, survey respondents are not always clear about how they are supposed to answer the questions and they draw on a range of question characteristics to figure out how to respond. These cues may include visual features of the question, such as its position on the screen. Tourangeau, Rips, and Rasinski (2000) list a number of potential reasons why respondents may misinterpret questions. To their list, we would add the following: Respondents may assign meanings to visual cues, such as screen position, that are not actually intended to convey any meaning.

As a practical matter, our results provide one more reason for presenting one item per screen in a Web survey, especially items that ask for evaluative ratings. Tourangeau and his coauthors (2004) found that putting items together on the same screen encouraged respondents to gloss over the differences among them. In these six new studies, putting two or more items on the same screen meant putting them at different vertical positions. The effects we observe may be small, but it is likely that similar screen-position effects would be observed in other modes of data collection that involve presenting items on a computer screen, such as audio computer-assisted self-interviewing (ACASI). It is also possible that similar effects would occur in paper questionnaires. The position effects we find could be considered a type of context effect. As with other context effects, the effects of the position of an item on the screen hardly render ratings data invalid (cf. Tourangeau, Singer, and Presser 2003). Our analysis of the cross-wave correlations in experiments 4 and 5 suggests that the screen position of an item does not have a consistent effect on the reliability of the answers.

Screen position effects may nonetheless jeopardize comparisons across surveys that present the same items but in different screen positions. Just as it is important to keep the context the same when comparing results across surveys, it is also important to hold the screen position constant. By the same token, it may be important to keep the position of an item within a grid the same when attempting to compare the results with an earlier study.

Supplementary Data

Supplementary data are freely available online at <http://poq.oxfordjournals.org/>.

References

- American Association for Public Opinion Research. 2011. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. 7th ed. Deerfield, IL: AAPOR.
- Barsalou, Lawrence W. 2008. "Grounded Cognition." *Annual Review of Psychology* 59:617–645.
- Carbonell, Jaime G. 1983. "Derivational Analogy in Problem Solving and Knowledge Acquisition." In *Proceedings of the International Machine Learning Workshop*, edited by R. S. Michalski, 12–18. Urbana: Department of Computer Science, University of Illinois at Urbana-Champaign.
- Couper, Mick P., Michael W. Traugott, and Mark J. Lamias. 2001. "Web Survey Design and Administration." *Public Opinion Quarterly* 65:230–253.
- Lakoff, George, and Mark Johnson. 1980. *Metaphors We Live By*. Chicago: University of Chicago Press.
- Lee, Angela Y., and Aparna A. Labroo. 2004. "The Effect of Conceptual and Perceptual Fluency on Brand Evaluation." *Journal of Marketing Research* 41:151–165.
- Lipsey, Mark W., and David B. Wilson. 2001. *Practical Meta-Analysis*. Thousand Oaks, CA: Sage Publications.
- Meier, Brian P., and Michael D. Robinson. 2004. "Why the Sunny Side Is Up: Associations between Affect and Vertical Position." *Psychological Science* 15:243–247.
- O'Muircheartaigh, Colm, George Gaskell, and Daniel B. Wright. 1995. "Weighing Anchors: Verbal and Numeric Labels for Response Scales." *Journal of Official Statistics* 11:295–307.

- Pandelaere, Mario, Kobe Millet, and Bram Van den Bergh. 2010. "Madonna or Don McLean? The Effect of Order of Exposure on Relative Liking." *Journal of Consumer Psychology* 20:442–451.
- Reber, Rolf, Piotr Winkielman, and Norbert Schwarz. 1998. "Effects of Perceptual Fluency on Affective Judgments." *Psychological Science* 9:45–48.
- Sakshaug, Joe, Roger Tourangeau, Jon A. Krosnick, Allison Ackermann, Ari Malka, Matthew DeBell, and Chintan Turakhia. 2009. "Dispositions and Outcome Rates in the 'Face-to-Face/Internet Survey Platform' (the FFISP)." Paper presented at the Annual Meeting of the American Association of Public Opinion Research, Hollywood, FL, USA.
- Schwarz, Norbert. 1996. *Cognition and Communication: Judgmental Biases, Research Methods, and the Logic of Conversation*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Schwarz, Norbert, Carla E. Grayson, and Barbel Knäuper. 1998. "Formal Features of Rating Scales and the Interpretation of Question Meaning." *International Journal of Public Opinion Research* 10:177–183.
- Schwarz, Norbert, Barbel Knäuper, Hans J. Hippler, Elisabeth Noelle-Neumann, and Leslie Clark. 1991. "Rating Scales: Numeric Values May Change the Meaning of Scale Labels." *Public Opinion Quarterly* 55:570–582.
- Smith, Tom W. 1995. "Little Things Matter: A Sampler of How Differences in Questionnaire Format Can Affect Survey Responses." JSM Proceedings of the American Statistical Association, Survey Research Methods Section, 1046–1051.
- Toepoel, Vera, and Don A. Dillman. 2008. *Words, Numbers, and Visual Heuristics in Web Survey: Is There a Hierarchy of Importance?* Discussion Paper, Tilburg University.
- Tourangeau, Roger, Mick P. Couper, and Frederick G. Conrad. 2004. "Spacing, Position, and Order: Interpretive Heuristics for Visual Features of Survey Questions." *Public Opinion Quarterly* 68:368–393.
- . 2007. "Color, Labels, and Interpretive Heuristics for Response Scales." *Public Opinion Quarterly* 71:91–112.
- Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski. 2000. *The Psychology of Survey Response*. New York: Cambridge University Press.
- Tourangeau, Roger, Eleanor Singer, and Stanley Presser. 2003. "Context Effects in Attitude Surveys: Effects on Remote Items and Impact on Predictive Validity." *Sociological Methods and Research* 31:486–513.
- Wells, Gary L., and Richard E. Petty. 1980. "The Effects of Over Head Movements on Persuasion: Compatibility and Incompatibility of Responses." *Basic and Applied Social Psychology* 1:219–230.
- Wright, Daniel B., and Marianna E. Carlucci. 2011. "The Response Order Effect: People Believe the First Person Who Remembers an Event." *Psychonomic Bulletin & Review* 18:805–812.