Sociological Methods & Research

http://smr.sagepub.com

# Designing Scalar Questions for Web Surveys

Leah Melani Christian, Nicholas L. Parsons and Don A. Dillman

*Sociological Methods Research* 2009; 37; 393
DOI: 10.1177/0049124108330004

The online version of this article can be found at:
http://smr.sagepub.com/cgi/content/abstract/37/3/393

Published by:
SAGE
http://www.sagepublications.com

**Sociological Methods & Research**
Volume 37 Number 3
February 2009 393-425

10.1177/0049124108330004
http://smr.sagepub.com
hosted at
http://online.sagepub.com

# Designing Scalar Questions for Web Surveys

Leah Melani Christian
*Pew Research Center for the People & the Press, Washington, DC*
Nicholas L. Parsons
*Eastern Connecticut State University, Willimantic, CT*
Don A. Dillman
*Washington State University, Pullman, WA*

This paper explores how the visual design of scalar questions influences responses in web surveys. We present the results of five experiments embedded in two web surveys of university students. We find that consistently presenting the positive end of the scale first did not impact responses but increases response times. Displaying the categories in multiple columns influence how respondents process the scale and increase response times. Separating the midpoint, "don't know" option, or endpoints spatially does not impact responses when the visual and conceptual midpoint align. Removing the graphical layout of the scale influences responses when lower numbers indicate more positive categories and increases response time. Finally, response times are longer for polar point scales with numeric labels, but there are no differences in responses. Overall, our results suggest that the visual design of response scales impacts measurement, but that some manipulations produce larger and more significant differences than others.

***Keywords:*** *web surveys; visual design; scalar questions; measurement*

Web surveys offer opportunities and challenges for designing survey questions. The proliferation of Internet surveys over the past decade has resulted in heightened attention to the visual design of survey questions and new features that can be integrated into Web surveys. This raises an important question of how Web surveys measure respondent characteristics and opinions, particularly when it is more difficult to ensure that

respondents perceive questions similarly. Unlike paper questionnaires, for which, once printed, the overall design and layout does not change, a Web survey may appear differently on the respondent's computer screen depending on the type of computer and browser used as well as individual configurations and settings. This issue has become more important as evidence increasingly shows that small changes in how questions appear to respondents can influence reporting of the characteristics that researchers are trying to measure (Dillman, Smyth, and Christian 2009).

Responses to ordinal scale questions can be particularly affected by how the question and response categories are presented. Since respondents are asked to select the answer that best represents their attitude or behavior along an implied continuum often marked by vague quantifiers, respondents gain meaning from how the response categories are displayed (Schwarz 1996). In addition, new response formats, such as drop-down menus and visual analog scales, and graphics are increasingly used for scalar questions in Web surveys. Thus, Web surveyors need to pay particular attention to how response categories are labeled and the overall layout and design of the response scale to ensure that scalar questions are accurately measuring respondents' attitudes and behaviors.

Our purpose in this article is to discuss some ways in which the visual design of scalar questions influences responses in Web surveys. In addition to reviewing some of the previous literature on designing response scales for Web surveys, we present the results of five experiments, each with multiple replications, from two Web surveys of random samples of university undergraduate students. The experiments test the effect of (a) reversing the order of the response categories; (b) presenting the scale in a vertical or horizontal layout; (c) the spacing of the midpoint, endpoints, and *don't know* option; (d) removing the graphic representation of the scale entirely; and (e) providing numeric labels. Overall, we expect that manipulating the design of the response scale can have significant impacts on the measurement of opinions and attitudes in Web surveys.

**Authors' Note:** Analysis of these data was made possible by funds provided to the Washington State University Social and Economic Sciences Research Center (SESRC) under Cooperative Agreement #43-3AEU-1-80055 with the USDA-National Agricultural Statistics Service, supported by the National Science Foundation, Division of Science Resource Statistics. We would also like to thank Jolene D. Smyth, Thom Allen, and Bruce Austin for their contributions to the research reported in this paper.

# Theoretical Background

## The Web as a Survey Mode

Web surveys can be seen as a hybrid mode, sharing features with self-administered and interview-administered surveys. Web surveys are similar to paper surveys in that they are primarily self-administered; respondents control the delivery of the stimulus and choose when to begin and quit the survey, the pace at which they complete the survey, and how they process each of the question screens. In addition, Web surveys rely primarily on visual communication, where respondents perceive and process the information in stages using their visual sensory system (Dillman et al. 2009; Ware 2004).

In contrast to paper surveys, Web surveys can incorporate audio and video to more closely approximate an interview-administered survey. Web researchers have begun to experiment with offering pictures of interviewers' faces (often with different characteristics or facial expressions), audio where questions are read to respondents, and animated agents or video of interviewers that can try to help guide respondents as they complete the survey (Schober and Conrad 2007). However, introducing interviewers can also introduce biases; the race (Krysan and Couper 2003) and gender of the interviewer (Fuchs 2008) have been shown to influence responses in Web surveys.

Internet surveys can also integrate a number of interactive features that can guide respondents and provide customized feedback as they complete the survey (Couper 2005). The flexibility of computer programming allows respondents to be navigated to appropriate sections and follow-up questions that apply to them. In addition, tailored responses can also be provided based on information that respondents submit or that is already known from the sample data or previous surveys. Thus, Web surveys have the potential to better engage respondents and customize the questionnaire than is possible in paper self-administered surveys. Additionally, respondents on the Web maintain more control over how they complete the survey than in interview-administered surveys.

Web surveys also face a new challenge—the influence of technological capabilities on how respondents experience the survey—that sets this mode apart from mail, face-to-face, and telephone surveys. Although computers have been used for decades in computer-assisted telephone (CATI) and personal interviewing (CAPI) as well as by respondents in computer-assisted self-administered interviewing (CASI), the surveyor usually has

direct control over nearly all aspects of the computers used. However, respondents usually complete Web surveys on their own computers that are connected to the Internet, which can result in substantial differences in how respondents experience the survey.

There is a great deal of variation in computer configurations, including the platform or operating system, Internet browser, speed of the processor, amount of memory and hard drive capabilities, and a variety of individual user settings as well as differences between desktop and laptop computers (e.g., screen size and input mechanisms). In addition, respondents must have access to the Internet, and connection speeds may vary greatly depending on the population surveyed. Last, knowledge of computer technology and how to navigate the Internet may differ for respondents, particularly by age. All of these factors can influence how respondents interact with the Web survey and the answers they provide.

In addition to the variation on the respondents' side, surveyors also differ widely in the technology they use to construct and field Web surveys. Web survey designers need computer hardware, a hosting server, and specific software to effectively design and conduct Web surveys. Variation in the features offered by different Web design software may limit (or enhance) the capabilities available to survey designers. In addition, the technology and software many survey designers use can often be far more advanced than what most respondents may have available to them (Dillman et al. 2009). Taken together, all of these factors may result in the Web survey designer seeing the survey questions quite differently from how the respondent sees them. Thus, Web surveys, specifically the visual presentation of Web surveys, is influenced both at the design and answering stages of the survey process by the technological resources and capabilities of the survey designers, administrators, and respondents. This is of particular concern as growing literature suggests that in addition to the words used to pose questions, the visual presentation of survey questions influences how people respond (for a summary, see Dillman et al. 2009).

## The Web and Visual Design

Survey researchers first began exploring how visual design affected respondents as they navigated paper questionnaires and responded to individual questions (Christian and Dillman 2004; Jenkins and Dillman 1997; Redline and Dillman 2002; Redline et al. 2003). However, the rise of Web surveys over the last decade has heightened interest in visual design effects and encouraged research and experimentation on how specific visual

manipulations affect respondents as they complete Web questionnaires. Since respondents to paper and most Web surveys lack interviewers to help guide them through the survey, the design of the questionnaire is extremely important in obtaining unbiased answers (Schwarz 1996).

Visual design influences not only how respondents organize the information on the screen or page and understand the general layout, but also how they process each individual question and its component parts (Dillman et al. 2009). Upon receiving a questionnaire, respondents first quickly scan the page and use bottom-up processing to get a sense of what is required of them; perception in this stage is focused only on the visual information presented. Respondents then begin to focus on the task of responding to the question, where visual perception is influenced by the context of the situation and the respondent's knowledge, expectations, and prior experiences. Thus, when the visual layout of the questions is not consistent with past experiences and expectations, respondents may experience confusion, provide incorrect or unintended answers, and take more time to respond. Because of this, poor visual design can also result in more respondents terminating the survey before they finish, thereby contributing to nonresponse error.

In addition to understanding the words on the page, respondents gain meaning from other visual elements, such as numbers, symbols, and graphics, which can affect how they comprehend what is being asked. The way the words and other visual elements are presented also influences how respondents perceive them; their properties, such as the element's size, shape, color, or location, can be used to increase or decrease the amount of attention respondents give each element (Dillman et al. 2009). For example, respondents may notice elements that "pop out" from the other elements on the page because of their visual differences.

Web surveying has allowed researchers to more comprehensively understand the effects of visual design by analyzing client-side paradata. Paradata can be collected while respondents complete the survey and can record the length of time between when the screen is loaded and when the respondent submits an answer, the order in which respondents select answers or complete questions, and whether they change their answer before selecting a final response (Heerwegh 2003). Response time can be one indicator of how people are responding to the visual design of particular questions (Stern 2008). Long response times may indicate that the visual design of a particular question makes cognitive processing of that question difficult for respondents.

## Designing Scalar Questions

There has been a great deal of research demonstrating how responses to scalar questions can be affected by the way the question and response scale are presented and the mode used to survey respondents (for a summary, see Dillman et al. 2009). In particular, previous research suggests that respondents gain information not only from the verbal labels assigned to each category but also from a category's position relative to other categories. The presentation of ordinal scales is particularly important since respondents use the answer categories to help understand what the question is asking and how to report their response.

When designing scalar questions, there are a variety of design choices that can affect how well the scale measures the respondent's underlying opinion or behavior. Researchers must decide how many categories to provide, whether to verbally and numerically label all or some of the categories, whether to present the scale visually, in what order to present the categories, whether to display them vertically or horizontally, whether to display them an equal distance from one another, and other design choices depending on the type of scale. Web surveyors may also choose to unfold longer scales by offering only a few response options at first (e.g., categories on each end of the scale and the midpoint) and then follow up with additional categories based on respondents' answers. In addition, images, such as logotypes or facial expressions, can be used to label the response categories.

New response formats are also available on the Web for asking scalar questions. A drop-down box can be used, where the response options are revealed once respondents click on the box, which can reduce download times and save screen space compared to displaying all of the options at once. However, research has shown that some people find them more difficult to use (Heerwegh and Loosveldt 2002) and that it is important not to display any of the options in the box until respondents click on it; otherwise, the visible options may be selected more often (Couper et al. 2004). Visual analog scales can also be used where respondents slide a mark along the scale line to select where they fit on the continuum (the computer can record the exact number, including decimals, to represent each respondent's answer). Research on visual analog scales suggests that although response times are longer for these types of scales, response distributions are similar to scales using radio buttons, where the options are presented horizontally (Couper et al. 2006) or vertically (Thomas and Couper 2007).

*Category order.* A great deal of research has focused on the order of response categories in nominal and ordinal scale questions. It has been argued that in self-administered surveys, respondents are more prone to select early listed categories (a primacy effect), and in interview surveys, to select from among the categories listed last (a recency effect). Krosnick and Alwin (1987) found support for these predicted patterns of primacy and recency effects when they compared responses to questions where the answer categories were presented verbally in interviews to answers where respondents read the choices from a show card. However, the evidence for widespread effects of this nature in telephone and mail surveys is quite mixed (Dillman et al. 1996; Dillman et al. 1995; Stern, Dillman, and Smyth 2007).

Research on the Web has shown some evidence of primacy effects. In an analysis of response order effects in check-all and forced-choice questions, Smyth et al. (2006) found evidence of primacy effects in check-all-that-apply nominal scale questions among those respondents who spent less than the mean amount of time responding to the question. Couper et al. (2004) found some evidence of primacy effects in a Web survey when testing response order effects using a radio button format and a drop-down box. Drawing on the original work of Krosnick and Alwin (1987), they proposed a cognitive explanation for the response order effects they found: Response options presented earlier were subject to deeper cognitive processing, which increased the likelihood of their selection. Mahon-Haft and Dillman (2007) compared a variety of scales and found some evidence of primacy, but overall the results were inconclusive. Similarly, Christian, Dillman, and Smyth (2007a) found that presenting the positive or negative category first did not affect responses on the Web.

*Scale layout.* In addition to the order in which categories are presented, the overall layout of the scale can affect responses. Smith (1995) found that the graphic layout of a socioeconomic status scale affected responses in an international survey. In all countries except the Netherlands, the results reflected existing information about stratification within these countries. Respondents in the Netherlands responded to a different version of the scale than respondents in other countries; they received a pyramid version in which the lower response boxes were wider than those on the middle and top, while respondents in other countries received a ladder version in which equally sized boxes were stacked upon one another. Smith speculated that respondents to the pyramid version were more likely to place themselves lower on the scale, assuming that the researcher's knowledge of the actual stratification in Dutch society meant that the average was

closer to the bottom. Schwarz, Grayson, and Knäuper (1998) confirmed Smith's speculations by conducting an experiment with university students asking them to subjectively rank their academic performance (using the ladder, pyramid, and onion format that was wider in the middle and narrow at the top). They found that the pyramid version produced lower academic performance scores than the ladder and onion formats. Both of these experiments show that the overall graphic layout of response scales influences the answers that respondents provide.

In previous research conducted using paper questionnaires, Christian and Dillman (2004) found that a linear scale layout, where all options were presented vertically, produced different answers than a nonlinear layout, where options were presented in two or three columns horizontally across the page. Respondents were more likely to select the third option in the nonlinear layout, where it appeared in the first row of the second column, than in the linear layout. More recently, similar results have been produced by Toepoel (2008), based on experiments in Dutch panel surveys of the general public.

Tourangeau, Couper, and Conrad argue "that respondents follow simple heuristics in interpreting the visual features of questions" (2004:370). They discuss five heuristics and report that middle means typical, left and top mean first, near means related, up means good, and like means close. They have reported the results from several experiments that test some these heuristics.

The "middle means typical" heuristic posits that respondents use the visual midpoint in a response scale as an "anchor or reference point for judging their own position" (Tourangeau et al. 2004:370). They found that when the visual midpoint of the scale does not align with the conceptual midpoint, responses are closer to the visual midpoint. Specifically, separating the *don't know* and *no opinion* categories from the other five substantive categories (ranging from *far too much* to *far too little*) in a vertical scale using either a divider line or additional spacing resulted in more people endorsing these categories. It also resulted in more endorsement of the response options that appear toward the top of the scale. However, when these options are not separated, respondents were more likely to endorse options that appear lower on the scale because the visual midpoint aligns with the category just below the midpoint. Similarly, when the conceptual midpoint was shifted to the left in a horizontal scale such that the visual midpoint was aligned more to the right, responses were more toward the right end of the scale.

According to the "left and top mean first" heuristic, respondents expect the first item in a response scale to represent the most positive or negative

option. Tourangeau et al. (2004) found that response times were significantly shorter for three of the six items when the scale was consistent with this heuristic (the scale progressed in logical order) than when the midpoint was last or when the categories were not ordered at all. The differences were stronger for the two attitudinal questions (using agree-disagree scales) than the four behavioral frequency items. Tourangeau et al. (2004) also proposed the "up means good" heuristic, although they did not test it, which suggests that respondents expect the most positive category to be presented first when the response categories are listed vertically. They posited that response times will be quicker and responses will be more reliable when the scale is presented consistently with this heuristic. However, presenting the positive category first may encourage primacy.

*Numeric labels and scale layout.* In a more recent article, Tourangeau, Couper, and Conrad (2007) reported experiments testing the "like means close" heuristic by analyzing how differential shading and numeric labels affect responses to scalar questions in Web surveys. They found a small effect of using different color hues for each end of the scale, but these effects disappeared when the scale included verbal or numeric labels for each category. Tourangeau et al. (2007) also replicated earlier findings by Schwarz et al. (1991) and found that fewer respondents select categories from the low end of the scale when they are labeled with negative numbers (−1 to −5) than with positive numbers (1 to 5). Respondents interpret the endpoint label for the low end of the scale as more negative or extreme when negative numbers are used. They also found that polar point scales with numeric labels produced similar results to scales without numeric labels. Overall, they argue that there is "a hierarchy of features that respondents attend to, with verbal labels taking precedence over numerical labels and numerical labels taking precedence over purely visual cues like color" (Tourangeau et al. 2007:109).

Other researchers have tested the effects of removing the graphic layout of the scale entirely and having respondents write or type a number in an answer box corresponding to their response. This format is similar to how scalar questions are asked on the telephone, where a visual representation of the scale cannot be provided (Dillman et al. 2009). Christian and Dillman (2004) found that respondents to a paper survey provide more negative ratings when presented with a number box format than with a polar point scale, where the scale is graphically displayed and the endpoints have verbal and numeric labels. In another study conducted using paper surveys, Stern et al. (2007) also found that people provided more

negative responses and were more likely to select the *don't know* option when provided the number box version compared with a polar point scale and that these findings did not differ by age, education, or gender.

In both of these studies, higher numbers were assigned to negative categories and lower numbers to positive categories. In contrast, Christian et al. (2007a) found that there are no significant differences between a polar point–labeled scale and a number box version when higher numbers are assigned to the positive categories and lower numbers to the negative categories. They also found that responses to a number box version of the scale were significantly more positive when 5 was assigned to the positive endpoint and 1 to the negative endpoint than when the positive endpoint was labeled 1.

Overall, the literature suggests that various aspects of how scales are presented can influence responses. Previous experiments demonstrate that not only do the verbal labels attached to each category affect how respondents interpret the scale, but also numeric labels and the amount of space between answer categories influence the meaning respondents attach to the verbal labels associated with each category.

We now turn to presenting the results of several experiments designed to test the effects of category order, scale layout, and numeric labels on responses to scalar questions. First, we present results from reversing the order of how the categories are presented. Then, we discuss whether a nonlinear layout, where responses are displayed in multiple columns across the page, influences how people process the scale compared to a linear layout, where all options are presented in one vertical column. We also explore how the spacing of individual categories (*don't know*, midpoint, and endpoints) may affect responses to these categories and overall responses. In addition, we also analyze how scale layout and numeric labels may work together to affect responses by discussing the results of a number box format, where respondents report a number corresponding to their response, compared with a polar point scale, where the endpoints are verbally labeled and all of the categories are numerically labeled. Last, we discuss results comparing a polar point scale with and without numeric labels.

## Procedures

These experimental comparisons were embedded in two Web surveys of undergraduate students at Washington State University (WSU). The surveys focused on evaluating the student experience at WSU's main

campus in Pullman and included various methodological experiments designed to analyze how the visual presentation of survey questions influences responses. The first Web survey was conducted from March 11 to May 9, 2003, using a randomly selected sample of undergraduate students enrolled at the Pullman campus. The response rate for this survey was 53 percent, with 1,591 (of 3,004) students responding. The second Web survey was conducted from November 4 to December 31, 2003. A total of 1,565 of the 3,043 students sampled completed the entire survey, for a response rate of 52 percent. Four versions of each Web survey were fielded; a random number generator assigned students to one of the four versions after they entered their access code.

For both Web surveys, all of the students were initially contacted using postal mail and were provided with a US$2 incentive and an individual access code to gain entrance to the survey. We also sent an initial e-mail contact providing a link to the survey and the student's access code to all of those who had e-mail addresses on file (about two thirds of the sample). Subsequent contacts to nonrespondents were sent using postal mail and e-mail.

The overall design of both Web surveys was similar. All of the questions discussed here appeared on their own page in black text against a colored background with white answer spaces to provide contrast and better visibility. The question screens were constructed using HTML tables where proportional widths were programmed in order to maintain a consistent visual stimulus regardless of individual user screens. Cascading Style Sheets were used to automatically adjust font size and style to accommodate varying user screen resolutions. The first Web survey contained 21 questions and the second 25 questions.

To compare differences between response formats, we present results from *t* tests of means and chi-square tests. One-sided *t* tests of means are shown for differences when we suggest a hypothesis; otherwise, *t* tests are two-sided. We report the mean response time for each question and treatment and *t* tests comparing the mean response time for different treatments. The response time is the length of time in seconds between when the screen loads and when the respondent provides an answer. Outliers were removed at $\pm 2$ standard deviations from the mean. Since participants were randomly assigned to one of four versions, statistically significant differences in response times can be attributed to question effects, independent of any variation in individual respondent characteristics, such as attitude accessibility or cognitive abilities (Johnson 2004).

# Results

## Reversing the Order of the Response Categories

On the second Web survey, respondents were presented with 10 questions testing how the order of the response categories affects responses. In one version, the scale was presented with the most positive answer choice first (e.g., *very satisfied*). In the second version, the most negative answer choice was presented first (see Figure 1a). Because previous research suggests that switching the order of categories within the survey can lead to measurement error (Israel and Taylor 1990), respondents received all 10 questions with either the positive end or the negative end first. The literature on primacy would suggest that respondents are more likely to select the items presented first. In addition, the "up means good" heuristic discussed by Tourangeau et al. (2004) indicates that respondents expect the positive end of the scale to be presented first and may take longer to respond to the version with the negative responses first because of the additional cognitive requirements needed to process information that is inconsistent with a priori expectations.

We find that a higher proportion of respondents choose the most positive response option when it is presented first in only 3 of the experiments, but in 7 of the 10 experiments, more respondents select the extreme negative response option when it is first (see Table 1). Only three of the differences in the endpoints and overall response distributions are significant (Questions 4, 5, and 18). Mean responses to the version in which positive options were presented first are slightly lower for 7 of the 10 questions; however, only one difference of means test is statistically significant (Question 5). Significant differences in responses were found only for questions that listed options vertically rather than horizontally.

Paradata results indicate that response times are longer when the negative, rather than the positive, end of the scale is presented first in 8 of the 10 questions. This finding supports the hypothesis that respondents take longer to cognitively process information when it is presented in a format inconsistent with their expectations. Because the responses overall tended to be more positive, the longer response times may also reflect that respondents to the version with the negative categories first had to process more categories before finding an appropriate response. In addition, the last 2 questions did not produce statistically significant time differences, which suggests that respondents who received the version with response categories presented in reverse order eventually learned to cognitively process the information and

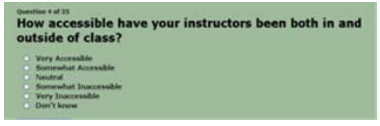
*(text continues on p. 409)*

## Figure 1
## Examples of Experimental Comparisons

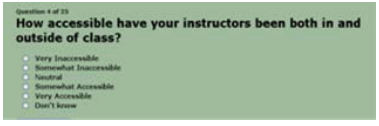
**a. Positive vs. negative response options first**

*10 comparisons*
7 satisfied/dissatisfied
1 accessible/inaccessible
1 desirable/undesirable
1 excellent/terrible

**Positive options first**



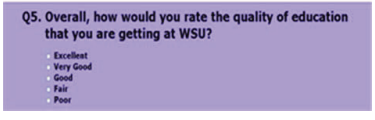

**Negative options first**



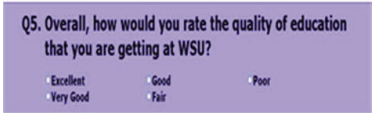

**b. Linear vs. nonlinear**

*4 comparisons*
3 excellent/poor
1 accessible/inaccessible

**Linear**



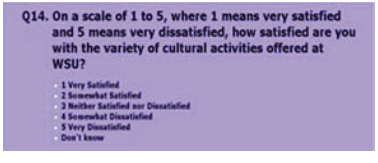

**Nonlinear**



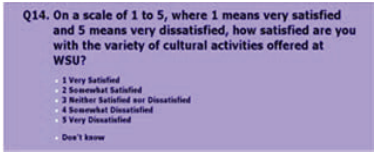

**c. Spacing of "Don't Know"**

*2 comparisons*
2 satisfied/dissatisfied

**Evenly spaced**




**Space before Don't Know**



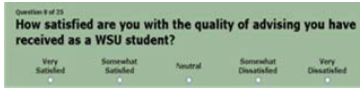

*(continued)*

# Figure 1
## (continued)

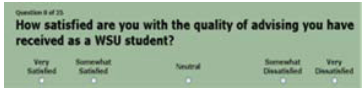
**d. Unequal spacing of midpoint and endpoints**

*7 comparisons*
5 satisfied/dissatisfied
1 desirable/undesirable
1 excellent/terrible

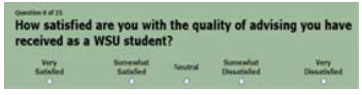
**Equally spaced**




**Isolated midpoint**



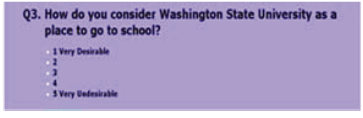

**Isolated endpoints**



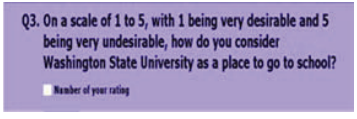

**e. Polar point vs. number box**

*6 comparisons*
4 satisfied/dissatisfied
1 desirable/undesirable
1 excellent/terrible

**Polar point with numbers**




**Number box**




**f. Polar point with vs. without numbers**

*4 comparisons*
2 satisfied/dissatisfied
1 desirable/undesirable
1 outstanding/terrible

**Polar point without numbers**

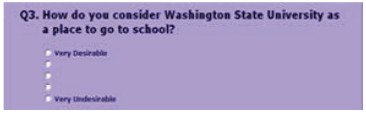


**Table 1**
**Positive Versus Negative End of Response Scale Presented First: Results From Web Survey 2**

| Q | Category Presented First | *N* | Percentage of Respondents Selecting Category | | | | | | Overall Chi-Square | | Chi-Square Endpoints | | Mean | Difference of Means | | Mean Time | Difference of Means | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1+ | 2 | 3 | 4 | 5– | DK | $\chi^2$ | *p* | $\chi^2$ | *p* | | *t* test | *p* | | *t* test | *p* |
| 2[a] | Positive | 413 | 50.6 | 38.0 | 5.8 | 5.1 | 0.5 | | 6.25 | .181 | 4.43 | .112 | 1.67 | –1.26 | .104 | 10.72 | –2.24 | .013 |
| | Negative | 428 | 50.9 | 34.6 | 5.4 | 7.0 | 2.1 | | | | | | 1.75 | | | 11.67 | | |
| 4[b] | Positive | 370 | 28.6 | 41.6 | 19.2 | 8.4 | 1.4 | 0.8 | 14.45 | .013 | 8.40 | .038 | 2.11 | –0.57 | .283 | 10.45 | –5.84 | .000 |
| | Negative | 425 | 32.2 | 35.5 | 14.4 | 12.7 | 2.6 | 2.6 | | | | | 2.16 | | | 13.11 | | |
| 5[c] | Positive | 393 | 29.0 | 40.2 | 22.2 | 6.6 | 2.0 | | 36.93 | .000 | 7.91 | .019 | 2.12 | –5.06 | .000 | 9.53 | –3.52 | .001 |
| | Negative | 425 | 21.9 | 26.6 | 36.7 | 10.6 | 4.2 | | | | | | 2.49 | | | 10.86 | | |
| 8[c] | Positive | 410 | 17.3 | 32.4 | 17.9 | 22.9 | 9.5 | | 3.00 | .885 | 1.08 | .576 | 2.75 | –0.54 | .725 | 9.66 | –2.19 | .015 |
| | Negative | 423 | 18.9 | 28.1 | 18.0 | 23.9 | 11.1 | | | | | | 2.80 | | | 10.37 | | |
| 10[d] | Positive | 386 | 17.6 | 17.1 | 18.4 | 9.1 | 10.1 | 27.7 | 5.85 | .321 | 4.29 | .233 | 2.68 | 0.98 | .840 | 10.92 | –3.11 | .001 |
| | Negative | 420 | 20.7 | 16.2 | 13.0 | 8.6 | 9.8 | 31.7 | | | | | 2.57 | | | 12.44 | | |
| 13[e] | Positive | 402 | 14.7 | 18.4 | 23.9 | 11.2 | 8.0 | 23.9 | 8.00 | .156 | 3.03 | .395 | 2.73 | 0.54 | .707 | 10.42 | –5.64 | .000 |
| | Negative | 418 | 12.2 | 24.2 | 19.6 | 12.9 | 5.7 | 25.4 | | | | | 2.68 | | | 12.40 | | |
| 17[a] | Positive | 395 | 31.6 | 42.8 | 13.5 | 9.6 | 2.5 | | 5.00 | .287 | 2.80 | .242 | 2.09 | –0.58 | .281 | 9.34 | –4.91 | .000 |
| | Negative | 417 | 33.1 | 38.1 | 16.0 | 8.2 | 4.6 | | | | | | 2.13 | | | 10.96 | | |

*(continued)*

**Table 1**
**(continued)**

| Q | Category Presented First | *N* | Percentage of Respondents Selecting Category | | | | | | Overall Chi-Square | | Chi-Square for Endpoints | | Mean | Difference of Means | | Mean Time | Difference of Means | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1 + | 2 | 3 | 4 | 5– | DK | $\chi^2$ | *p* | $\chi^2$ | *p* | | *t* test | *p* | | *t* test | *p* |
| 18[f] | Positive | 401 | 20.9 | 34.9 | 25.8 | 14.2 | 4.2 | | 11.81 | .019 | 10.72 | .005 | 2.46 | –0.50 | .307 | 9.65 | –1.80 | .036 |
| | Negative | 416 | 24.0 | 31.3 | 24.7 | 10.6 | 9.4 | | | | | | 2.50 | | | 10.21 | | |
| 19[g] | Positive | 379 | 3.2 | 21.4 | 35.6 | 28.2 | 11.6 | | 0.61 | .962 | 0.21 | .909 | 3.24 | –0.11 | .457 | 9.81 | 0.16 | .564 |
| | Negative | 416 | 2.9 | 20.0 | 37.7 | 28.6 | 10.8 | | | | | | 3.24 | | | 9.76 | | |
| 25[h] | Positive | 368 | 50.3 | 32.1 | 10.8 | 5.7 | 1.1 | | 2.10 | .717 | 1.11 | .587 | 1.75 | –0.76 | .224 | 7.82 | –1.10 | .136 |
| | Negative | 397 | 49.9 | 29.5 | 12.8 | 5.8 | 2.0 | | | | | | 1.81 | | | 8.15 | | |

Note: Categories are listed vertically for Questions 2, 4, 5, 10, 13, 18, and 25; categories are listed horizontally for Questions 8, 17, and 19; there is an isolated midpoint on Questions 2, 5, 17, 18, and 25. DK = *don't know* or *does not apply to me*.

a. Category labels for Questions 2 and 17: 1 = *very satisfied*, 2 = *somewhat satisfied*, 3 = *neither satisfied nor dissatisfied*, 4 = *somewhat dissatisfied*, 5 = *very dissatisfied*.

b. Question 4: 1 = *very accessible*, 2 = *somewhat accessible*, 3 = *neither accessible nor inaccessible*, 4 = *somewhat inaccessible*, 5 = *very inaccessible, don't know*.

c. Questions 5 and 8: 1 = *very satisfied*, 2 = *somewhat satisfied*, 3 = *neutral*, 4 = *somewhat dissatisfied*, 5 = *very dissatisfied*.

d. Question 10: 1 = *very satisfied*, 5 = *very dissatisfied, does not apply to me*.

e. Question 13: 1 = *very satisfied*, 5 = *very dissatisfied, don't know*.

f. Question 18: 1 = *very desirable*, 5 = *very undesirable*.

g. Question 19: 1 = *excellent*, 2 = *good*, 3 = *fair*, 4 = *poor*, 5 = *terrible*.

h. Question 25: 1 = *very satisfied*, 5 = *very dissatisfied*.

respond in the same amount of time as those who received the version in which the most positive answer choice was listed first.

## Presenting the Scale in Linear Versus Nonlinear Format

Four questions from the first Web survey were designed to test response differences between a linear and nonlinear layout of the response scale. In the linear version, response options were presented vertically in a single column, beginning with the positive endpoint and ending with the negative endpoint. In the nonlinear version, the response options were in the same order but presented in two or three columns, depending on the length of the scale (see Figure 1b). In the linear version, we expect that respondents will process the options in the intended order from top to bottom. However, respondents to the nonlinear version may process the options in rows, reading horizontally from left to right in the first row and then the next row, or they may process the options in columns, moving vertically from top to bottom in the first column and then the next column. If some respondents are processing the scale horizontally, from left to right, we expect that more people will select the third category located to the right of the first category in the nonlinear version, as found in previous research using paper surveys (Christian and Dillman 2004). Response times should also be longer for the nonlinear format because respondents may have a more difficult time processing and comprehending the scale.

In three of the four questions, respondents who received the nonlinear layout of the scale were more likely to choose the third response option (see Table 2). On average, respondents to the nonlinear version provided more negative ratings than respondents to the linear version; three of the differences in the means are statistically significant. For all four questions, response times were significantly longer for the nonlinear version than the linear version, which indicates that the linear version may be easier for respondents to process.

## Spacing of the Individual Response Options

*Spacing of the don't know option.* We also tested how unequal spacing of individual answer categories affected responses. Two questions in the second Web survey compared the effect of adding an additional space before the *don't know* option (see Figure 1c). In one version, the *don't know* category was placed directly under the fifth option, making the visual midpoint appear between the midpoint and the fourth response

**Table 2**
**Linear Versus Nonlinear Layout: Results From Web Survey 1**

| | | | Percentage of Respondents Selecting Category | | | | | | Overall Chi-Square | | Mean | Difference of Means | | Mean | Difference of Means | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Q | Version | *N* | 1+ | 2 | 3 | 4 | 5– | DK | $\chi^2$ | *p* | | *t* test | *p* | Time | *t* test | *p* |
| 5[a] | Linear | 351 | 12.5 | 49.6 | 30.5 | 6.3 | 1.1 | | 6.37 | .173 | 2.34 | 1.86 | .031 | 6.22 | –3.27 | .001 |
| | Nonlinear | 438 | 9.8 | 45.4 | 35.4 | 8.9 | 0.5 | | | | 2.45 | | | 7.05 | | |
| 7[b] | Linear | 431 | 50.1 | 45.0 | 4.6 | 0.2 | | | 9.25 | .026 | 1.55 | 2.94 | .002 | 6.27 | –7.14 | .000 |
| | Nonlinear | 437 | 40.3 | 53.1 | 6.0 | 0.7 | | | | | 1.67 | | | 8.35 | | |
| 12[a] | Linear | 434 | 24.0 | 42.4 | 27.9 | 4.8 | 0.9 | | 16.99 | .002 | 2.16 | 4.03 | .000 | 5.32 | –3.85 | .000 |
| | Nonlinear | 438 | 16.4 | 39.3 | 32.7 | 10.1 | 1.6 | | | | 2.41 | | | 6.25 | | |
| 15[c] | Linear | 431 | 13.9 | 39.4 | 15.6 | 1.2 | | 29.9 | 3.92 | .505 | 2.06 | 1.25 | .107 | 5.95 | –2.21 | .014 |
| | Nonlinear | 437 | 10.3 | 41.4 | 15.3 | 1.8 | | 31.1 | | | 2.13 | | | 6.48 | | |

Note: DK = *don't know*.
a. Category labels for Questions 5 and 12: 1 = *excellent*, 2 = *very good*, 3 = *good*, 4 = *fair*, 5 = *poor*.
b. Question 7: 1 = *very accessible*, 2 = *somewhat accessible*, 3 = *somewhat inaccessible*, 4 = *very inaccessible*.
c. Question 15: 1 = *excellent*, 2 = *good*, 3 = *fair*, 4 = *poor*, *don't know*.

option (*somewhat dissatisfied*). In the second version, the *don't know* option was separated from the substantive options by additional space, allowing the visual midpoint to align with the conceptual midpoint (*neither satisfied nor dissatisfied*). Both versions were fully labeled with a *don't know* category and five substantive options (i.e., *very satisfied* to *very dissatisfied*) numerically labeled from 1 to 5. According to the Gestalt principle of proximity, the addition of the space should encourage respondents to perceive the substantive options as one group and the *don't know* category separately, whereas respondents to the evenly spaced version should group all six options together (Dillman et al. 2009).

If respondents interpret the visual midpoint of the response options to be the conceptual midpoint, we expect higher means, indicating more responses lower on the scale when there is no additional space between the substantive options and the *don't know* category, as Tourangeau et al. (2004) showed. Separating the *don't know* option visually may encourage more people to select this category. Contrary to our hypothesis, mean responses were higher, representing more people selecting categories from the low end of the scale when the *don't know* option was separated from the other answer categories (see Table 3). In addition, respondents were no more likely to select the *don't know* option when it was visually separated than when it was not. However, we found that slightly more respondents selected the midpoint when the visual and conceptual midpoints aligned, a finding consistent with Tourangeau et al. Finally, paradata results comparing mean response times show that respondents took significantly longer to submit an answer when the *don't know* option was visually separated with space, which suggests that it took longer for respondents to perceive all of the options and formulate a response.

There are several reasons why our results may differ from those reported by Tourangeau et al. (2004). First, they included two nonsubstantive options (*don't know* and *no opinion*), which resulted in their visual midpoint aligning fully with the category just below the midpoint. Second, the scale labeling was quite different. Their scale ranged from *far too much* to *far too little*, whereas our scale ranged from *very satisfied* to *very dissatisfied*. We also included numeric labels for the substantive options, whereas their scale did not. Thus, even when the *don't know* was not separated by an additional space, it was distinguished from the other categories by its lack of a numeric label.

**Table 3**
**Extra Space or No Space Before *Don't Know* (DK) Category: Results From Web Survey 1**

| | | | Percentage of Respondents Selecting Category | | | | | | Overall Chi-Square | | Chi-Square for Midpoint | | | Difference of Means | | Mean | Difference of Means | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Q | Version | *N* | 1+ | 2 | 3 | 4 | 5– | DK | $\chi^2$ | *p* | $\chi^2$ | *p* | Mean | *t* test | *p* | Time | *t* test | *p* |
| 14 | Space | 348 | 14.4 | 22.7 | 35.9 | 7.8 | 3.5 | 15.8 | 4.59 | .468 | 3.56 | .059 | 2.56 | 1.30 | .097 | 12.07 | –4.06 | .000 |
| | No space | 433 | 15.9 | 26.1 | 29.6 | 8.3 | 2.5 | 17.6 | | | | | 2.46 | | | 10.24 | | |
| 17 | Space | 350 | 59.1 | 23.1 | 8.0 | 3.4 | 3.1 | 3.1 | 10.72 | .026 | 6.76 | .000 | 1.64 | 2.89 | .002 | 8.68 | –3.60 | .000 |
| | No space | 433 | 68.4 | 19.2 | 3.7 | 3.2 | 1.6 | 3.9 | | | | | 1.44 | | | 7.57 | | |

Note: Category labels are 1 = *very satisfied*, 2 = *somewhat satisfied*, 3 = *neither satisfied nor dissatisfied*, 4 = *somewhat dissatisfied*, 5 = *very dissatisfied, DK = don't know*.

*Unequal spacing of midpoint and endpoints.* We also tested how unequal spacing of the midpoint and endpoints affected responses. In the second Web survey, we compare a response scale where the five categories are evenly spaced to versions where the midpoint and endpoints are unequally spaced (see Figure 1d). In the unequally spaced midpoint version, there is more space between the midpoint and the two categories on each side, whereas the version with unequally spaced endpoints has more space between the endpoints and the middle three categories. In all three versions, the conceptual and visual midpoints align.

Isolating the midpoint or endpoints should serve to illuminate them when respondents first look at the scale, potentially increasing the number who may select that option. The principle of proximity also suggests that respondents will perceive the options that are closer together as a group (e.g., group the middle categories separately from each endpoint or the midpoint separately from the other categories on either side). The unequally spaced versions may take longer for respondents to process than when the options are evenly spaced. Overall, our results comparing the evenly spaced version to the unequally spaced midpoint and endpoints versions indicate very little differences in responses; none are statistically significant, and they yield no consistent pattern (see Table 4). In addition, response time results show significant differences for 3 of the 11 total comparisons, but these too yield no consistent pattern (the times for the unequally spaced versions are not consistently longer).

## Removing the Graphic Layout of the Scale

In both surveys, we tested the effects of removing the graphic representation of the scale entirely and asking people to respond by typing in a number corresponding to their response. The number box format is closer to how some scalar questions are asked on the telephone, where a visual display of the scale cannot easily be provided. The polar point scale graphically displays all of the response categories, providing additional information for respondents to use at the time they select their answer (see Figure 1e). In the scalar questions used in this experiment and the others reported here, lower numbers (i.e., 1, 2) represent more positive attitudes (e.g., *very satisfied*, *somewhat satisfied*), while higher numbers (i.e., 4, 5) represent more negative attitudes (e.g., *dissatisfied*, *very dissatisfied*). Similar to the previous findings, we expect that the number box format may yield more negative ratings. We also expect that response times will be higher for the number box version because (a) the complete removal of

**Table 4**
**Equal Versus Unequal Spacing of Midpoints and Endpoints: Results From Web Survey 2**

| | | | Percentage of Respondents Selecting Category | | | | | Overall Chi-Square | | | Difference of Means | | | Difference of Means | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Q | Version | *N* | 1+ | 2 | 3 | 4 | 5– | $\chi^2$ | *p* | Mean | *t* test | *p* | Mean Time | *t* test | *p* |
| 2[a] | Equal | 370 | 50.3 | 37.8 | 5.4 | 6.0 | 0.5 | 0.34 | .987 | 1.69 | 0.30 | .765 | 10.06 | 1.62 | .107 |
| | Isolated midpoint | 413 | 50.6 | 38.0 | 5.8 | 5.1 | 0.5 | | | 1.67 | | | 10.72 | | |
| 5[b] | Equal | 368 | 28.3 | 38.0 | 22.8 | 7.6 | 3.3 | 1.65 | .799 | 2.20 | 0.97 | .331 | 9.74 | 0.61 | .541 |
| | Isolated midpoint | 393 | 29.0 | 40.2 | 22.1 | 6.6 | 2.0 | | | 2.12 | | | 9.53 | | |
| 8[b] | Equal | 368 | 15.5 | 36.7 | 16.8 | 22.8 | 8.2 | 1.91 | .751 | 2.71 | –0.38 | .700 | 9.97 | 1.05 | .296 |
| | Isolated midpoint | 410 | 17.3 | 32.4 | 17.8 | 22.9 | 9.5 | | | 2.75 | | | 9.66 | | |
| 17[a] | Equal | 362 | 34.0 | 35.4 | 15.2 | 12.2 | 3.3 | 4.90 | .297 | 2.15 | 0.88 | .382 | 10.42 | 3.25 | .001 |
| | Isolated midpoint | 395 | 31.6 | 42.8 | 13.4 | 9.6 | 2.5 | | | 2.09 | | | 9.34 | | |
| 18[c] | Equal | 364 | 26.1 | 30.5 | 23.9 | 12.4 | 7.1 | 6.90 | .141 | 2.44 | –0.23 | .818 | 10.03 | 1.28 | .202 |
| | Isolated midpoint | 401 | 20.9 | 34.9 | 25.7 | 14.2 | 4.2 | | | 2.46 | | | 9.65 | | |
| 19[d] | Equal | 364 | 1.6 | 20.1 | 31.0 | 34.9 | 12.4 | 2.39 | .664 | 3.36 | 0.40 | .689 | 9.17 | 1.50 | .134 |
| | Isolated midpoint | 398 | 1.0 | 20.4 | 35.2 | 31.2 | 12.3 | | | 3.33 | | | 8.74 | | |
| 25[e] | Equal | 405 | 53.3 | 27.4 | 11.4 | 6.2 | 1.7 | 2.43 | .657 | 1.76 | 0.04 | .968 | 7.22 | 2.16 | .031 |
| | Isolated midpoint | 368 | 50.3 | 32.1 | 10.9 | 5.7 | 1.1 | | | 1.75 | | | 7.82 | | |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8[b] | Equal | 368 | 15.5 | 36.7 | 16.8 | 22.8 | 8.2 | 7.03 | .134 | 2.71 | 0.68 | .498 | 9.97 | 0.39 | .697 |
| | Isolated endpoints | 389 | 22.1 | 31.1 | 15.9 | 21.1 | 9.8 | | | 2.65 | | | 10.11 | | |
| 17[a] | Equal | 362 | 34.0 | 35.4 | 15.2 | 12.2 | 3.3 | 5.47 | .242 | 2.15 | 1.02 | .309 | 10.42 | 0.09 | .930 |
| | Isolated endpoints | 381 | 32.8 | 42.3 | 13.1 | 8.4 | 3.4 | | | 2.07 | | | 10.45 | | |
| 8[b] | Isolated midpoint | 410 | 17.3 | 32.4 | 17.8 | 22.9 | 9.5 | 3.18 | .529 | 2.75 | 1.06 | .289 | 9.66 | 1.36 | .173 |
| | Isolated endpoints | 389 | 22.1 | 31.1 | 15.9 | 21.1 | 9.8 | | | 2.65 | | | 10.11 | | |
| 17[a] | Isolated midpoint | 395 | 31.6 | 42.8 | 13.4 | 9.6 | 2.5 | 0.93 | .920 | 2.09 | 0.17 | .866 | 9.34 | 3.42 | .001 |
| | Isolated endpoints | 381 | 32.8 | 42.3 | 13.1 | 8.4 | 3.4 | | | 2.07 | | | 10.45 | | |

Note: Categories are listed vertically for Questions 2, 5, 18, and 19; categories are listed horizontally for Questions 8, 17, and 25.

a. Category labels for Questions 2 and 17: 1 = *very satisfied*, 2 = *somewhat satisfied*, 3 = *neither satisfied nor dissatisfied*, 4 = *somewhat dissatisfied*, 5 = *very dissatisfied.*

b. Questions 5 and 8: 1 = *very satisfied*, 2 = *somewhat satisfied*, 3 = *neutral*, 4 = *somewhat dissatisfied*, 5 = *very dissatisfied*.

c. Question 18: 1 = *very desirable*, 5 = *very undesirable*.

d. Question 19: 1 = *excellent*, 2 = *good*, 3 = *fair*, 4 = *poor*, 5 = *terrible*.

e. Question 25: 1 = *very satisfied*, 5 = *very dissatisfied.*

the answer scale eliminates the visual cues respondents use to help them comprehend the question and formulate an answer and (b) the response task requires respondents to translate their judgment to a number and enter that number into the box.

Consistent with our hypothesis, we find that the number box version produced significantly higher means, indicating more negative ratings, in five of the six comparisons (see Table 5). Chi-square tests were statistically significant for five of six questions. Overall, the larger means produced by the number box version may be a result of respondents associating higher numbers with more positive ratings. Previous research has shown that respondents tend to associate higher numbers with more positive categories and lower numbers with more negative categories and that responses may be negatively affected when lower numbers are assigned to more positive categories (Christian and Dillman 2004; Christian et al. 2007a).

As expected, respondents took significantly longer to answer the number box version than the polar point version in all six comparisons (see Table 5). The longer response time is probably attributable to both the difficulty in responding without a graphic representation of the scale and the additional steps required for submitting a response.

## Including or Excluding Numeric Labels on Polar Point Scales

Four questions on the first Web survey were designed to test the effect of including numeric labels by comparing a version of a polar point scale without numeric labels to one with numbers (see Figure 1f). Since the numeric labels are an additional source of information for respondents to process, we expect this supplementary information to influence the responses people provide and increase the time it takes to respond. With the exception of one question, mean responses are virtually identical regardless of whether numeric labels were included, and the response distributions are not significantly different between the two versions (see Table 6). Respondents took longer to submit an answer when provided with a numerically labeled polar point scale, which suggests that the addition of numeric language provides an additional source of information that respondents must cognitively process before submitting an answer. These results indicate that numeric labels may not be necessary when the visual layout itself, by equally spacing the response categories, conveys the measurement intent of the scale.

**Table 5**
**Polar Point Versus Number Box: Results From Web Surveys 1 and 2**

| | | | Percentage of Respondents Selecting Category | | | | | | Overall Chi-Square | | | Difference of Means | | Mean | Difference of Means | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Q | Version | *N* | 1+ | 2 | 3 | 4 | 5– | DK | $\chi^2$ | *p* | Mean | *t* test | *p* | Time | *t* test | *p* |
| 3[a] | Polar point | 438 | 24.7 | 43.2 | 24.4 | 6.4 | 1.4 | | 54.20 | .000 | 2.17 | –5.73 | .000 | 8.53 | –12.03 | .000 |
| | Number box | 350 | 19.1 | 36.3 | 17.7 | 18.9 | 8.0 | | | | 2.60 | | | 13.25 | | |
| 14[b] | Polar point | 436 | 13.8 | 20.6 | 25.2 | 7.6 | 3.7 | 29.1 | 46.83 | .000 | 2.53 | –4.93 | .000 | 10.63 | –11.65 | .000 |
| | Number box | 358 | 8.4 | 9.2 | 19.0 | 12.6 | 7.8 | 43.0 | | | 3.04 | | | 17.89 | | |
| 17[b] | Polar point | 438 | 60.5 | 21.5 | 10.1 | 1.6 | 1.1 | 5.3 | 94.96 | .000 | 1.54 | –8.13 | .000 | 7.66 | –10.50 | .000 |
| | Number box | 361 | 46.8 | 13.3 | 6.9 | 10.3 | 14.4 | 8.3 | | | 2.26 | | | 11.17 | | |
| 12[b] | Polar point | 402 | 13.7 | 37.6 | 29.6 | 7.7 | 0.5 | 11.0 | 34.19 | .000 | 2.37 | –5.64 | .000 | 13.41 | –4.49 | .000 |
| | Number box | 380 | 7.6 | 29.7 | 30.5 | 11.9 | 2.3 | 11.3 | | | 2.77 | | | 15.51 | | |
| 13[b] | Polar point | 401 | 14.5 | 18.5 | 23.9 | 11.2 | 7.9 | 23.9 | 18.40 | .002 | 2.73 | –3.46 | .000 | 10.42 | –3.25 | .001 |
| | Number box | 378 | 8.7 | 15.6 | 18.0 | 15.9 | 11.9 | 29.9 | | | 3.09 | | | 11.47 | | |
| 15[c] | Polar point | 403 | 5.0 | 20.6 | 31.0 | 21.8 | 9.4 | 12.2 | 3.28 | .657 | 3.11 | 0.03 | .514 | 11.45 | –4.97 | .000 |
| | Number box | 377 | 6.6 | 21.5 | 28.7 | 21.8 | 11.7 | 9.8 | | | 3.11 | | | 13.19 | | |

Note: DK = *don't know*. Questions 3, 14, and 17 are from Web Survey 1. Questions 12, 13, and 15 are from Web Survey 2.
a. Category labels for Question 3: 1 = *very desirable*, 5 = *very undesirable*.
b. Questions 14, 17, 12, and 13: 1 = *very satisfied*, 5 = *very dissatisfied*, *don't know*.
c. Question 15: 1 = *excellent*, 5 = *terrible*, *don't know*.

**Table 6**
**Polar Point With Versus Without Numbers: Results From Web Survey 1**

| | | | Percentage of Respondents Selecting Category | | | | | Overall Chi-Square | | | Difference of Means | | | Difference of Means | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Q | Version | *N* | 1+ | 2 | 3 | 4 | 5− | $\chi^2$ | *p* | Mean | *t* test | *p* | Mean Time | *t* test | *p* |
| 3[a] | With #s | 438 | 24.7 | 43.2 | 24.4 | 6.4 | 1.4 | 3.00 | .557 | 2.17 | 1.43 | .076 | 8.53 | 3.22 | .001 |
| | Without #s | 366 | 26.5 | 45.1 | 23.2 | 4.6 | 0.6 | | | 2.08 | | | 7.43 | | |
| 6[b] | With #s | 438 | 10.5 | 43.6 | 35.2 | 9.8 | 0.9 | 2.21 | .697 | 2.47 | 0.17 | .433 | 10.84 | 1.89 | .030 |
| | Without #s | 365 | 8.8 | 47.7 | 33.7 | 8.5 | 1.4 | | | 2.46 | | | 10.13 | | |
| 8[b] | With #s | 438 | 19.4 | 40.0 | 30.6 | 8.2 | 1.8 | 5.46 | .243 | 2.33 | 0.28 | .391 | 9.46 | 2.14 | .016 |
| | Without #s | 364 | 15.7 | 47.5 | 28.6 | 6.3 | 1.9 | | | 2.31 | | | 8.71 | | |
| 9[c] | With #s | 437 | 14.0 | 29.3 | 30.0 | 19.0 | 7.8 | 0.82 | .935 | 2.77 | 0.09 | .463 | 9.12 | 1.73 | .042 |
| | Without #s | 363 | 15.7 | 27.6 | 28.9 | 20.1 | 7.7 | | | 2.77 | | | 8.59 | | |

a. Category labels for Question 3: 1 = *very desirable*, 5 = *very undesirable*.
b. Questions 6 and 8: 1 = *very satisfied*, 5 = *very dissatisfied*.
c. Question 9: 1 = *outstanding*, 5 = *terrible*.

**Table 7**
**Summary of Experimental Comparisons and Findings**

| Experiment | | Significant Number of Comparisons by Total Number of Comparisons: Difference of Means *t* test | $\chi^2$ test | Different of Mean Response Time *t* test |
|---|---|---|---|---|
| Positive vs. negative end first | Table 1 | 1 of 10 | Overall 3 of 10<br>Endpoints 3 of 10 | 8 of 10 |
| Linear vs. nonlinear layout | Table 2 | 3 of 4 | 2 of 4 | 4 of 4 |
| Spacing of *don't know* | Table 3 | 1 of 2 | Overall 1 of 2<br>Midpoint 1 of 2 | 2 of 2 |
| Spacing of midpoints and endpoints | Table 4 | | | |
| Equal vs. isolated midpoint | | 0 of 7 | 0 of 7 | 2 of 7 |
| Equal vs. isolated poles | | 0 of 2 | 0 of 2 | 0 of 2 |
| Isolated poles vs. midpoint | | 0 of 2 | 0 of 2 | 1 of 2 |
| Polar point vs. number box | Table 5 | 5 of 6 | 5 of 6 | 6 of 6 |
| Polar point with vs. without numbers | Table 6 | 0 of 4 | 0 of 4 | 4 of 4 |

## Discussion and Conclusion

Results from past research and our experiments indicate that respondents often rely on more than just the question wording and verbal category labels when answering scalar questions. Specifically, respondents gain information from the overall layout of the scale, each category's physical position in relation to other responses on the scale, and any numeric labels that may be used. Similar to other research on visual design effects, our results indicate that some visual manipulations of the response scale produce large and significant differences and others more minor differences (see Table 7 for a summary of our results).

Overall, we find that consistently presenting the positive end of the scale first for all questions did not affect people's responses, but it did result in respondents providing a response more quickly. When the options are presented consistently with the "up means good" heuristic (Tourangeau et al. 2004), it seems to take less time for respondents to perceive and comprehend the scale, so they can provide a response faster. In addition, presenting the options linearly in one column facilitates respondents' processing of the scale and encourages them to process the categories in the same order, making it easier to provide a response. In contrast, presenting the categories in multiple columns different responses based on whether results in respondents process the options horizontally or vertically and in respondents taking longer to provide an answer. Our findings confirm that the midpoint is an important anchor for respondents when answering scalar questions. Separating the midpoint, *don't know* option, or endpoints through the use of space does not seem to affect responses as long as the visual and conceptual midpoints of the scale are aligned.

Additionally, we find that removing the graphic layout of the scale entirely and providing only a verbal description of the scale in the question stem significantly influence respondents' answers and increase response time when lower numbers indicate more positive categories. However, previous research has shown that when the numbers are assigned consistently with respondents' expectations, with higher number assigned to positive categories, removing the graphic display of the scale does not significantly affect responses (Christian et al. 2007a). Similarly, past research has shown that the use of negative numbers dramatically affects responses compared to the use of only positive numbers (Schwarz et al. 1991; Tourangeau et al. 2007). Finally, our results demonstrate that respondents take longer to respond to polar point scales with numeric labels because of the additional information they must cognitively process; however, there are no statistically significant differences in responses.

Further research on the visual design of scalar questions using general population samples within the United States and other countries is also needed. The generalizability of our results is limited because our sample included only undergraduate students, all from one university in the United States. Because they are enrolled in college and younger than the average population, students are more computer literate and probably find it easier to respond to surveys on the Internet. Although we tested our surveys on various types of computer platforms, browsers, and connection speeds, it is also likely that many of these university students completed the survey on computers and with connections that may be

better than what is found in the general population. Overall, future experiments should analyze how the visual design of scalar questions affects respondents differently, based on age, education, and other characteristics.

Traditionally, survey methodologists have focused on wording alone when constructing response scales, and the visual presentation of response scales has received little attention (Krosnick and Fabrigar 1997). This has sometimes led to the unfortunate practice of surveyors replicating the wording of questions from one study to another with the intent of comparing responses across studies, but changing the visual layout to fit with the style of a particular questionnaire design or because of space limitations. Now, based on a decade of research, as discussed in the theoretical background and the experiments reported in this article, it is clear that the visual attributes of scalar questions also influence answers and cannot be ignored.

As shown here and in previous research, not all differences in the visual presentation of scalar questions significantly influence respondents' answers. In addition, verbal and numeric labels may take precedence over subtle visual cues, particularly when respondents are focused on responding to individual questions (Tourangeau et al. 2007). However, it has also been demonstrated how visual, numeric, and symbolic information can be more powerful than verbal information (Christian, Dillman, and Smyth 2007b; Redline et al. 2003). In particular, visual information often takes precedence over verbal information when respondents first process the information presented (Ware 2004). The development of an overall theory of when and why visual layout makes a difference remains in its early stages (see Dillman et al. 2009). More research is needed to understand how people's answers are independently and jointly influenced by these features of survey design.

Our overall finding that the visual design of response scales makes a difference has significant practical implications. When designing response scales for paper and Web questionnaires, surveyors must not only consider how many categories to offer and what verbal labels to provide but also how the scale will be presented visually to respondents. It also seems important that when surveyors use questions from earlier surveys with the intent of comparing results, they attempt to maintain the visual qualities of previous questions as well as the question wording. Another implication is that at a time when mixed-mode surveys are being conducted with greater frequency, surveyors must recognize that attention needs to be given to maintaining similarity across survey modes. This challenge is especially

important for Internet surveys as they are often used in mixed-mode surveys, and those who respond by Web are often different from those who respond to another mode.

The mixing of aural and visual survey modes raises additional issues. Recent research has shown that respondents consistently provide more positive answers to scalar questions when asked by telephone than when surveyed by paper or Web (Christian 2007; Christian et al. 2007a; Dillman et al. forthcoming). Not only must attention be given to seeking solutions to these problems, but also it seems plausible that some visual layouts of questions are more likely to obtain similar results in aural modes than are others; this seems a particularly important issue for future research.

Finally, the results presented here provide insight into respondent burden. Traditionally, respondent burden has been thought of as an attribute of a question that was dependent on words and substance. The experimental data reported here show that even when people's responses are not affected, some visual layouts are more quickly processed than others (e.g., polar point scales without numbers). In other instances, not only does the visual design influence responses, but also it can affect the time taken to provide a response (e.g., polar point vs. number box when the more positive answers are assigned lower numbers). Response time is longer for formats that are difficult for respondents to process. It is increasingly clear that paradata have multiple uses for survey methodologists, by helping them to understand when questions are especially burdensome or misunderstood by respondents as well as the order in which respondents complete questions and select individual answers. As research focuses on what combination of words and visual design produce the best possible survey questions, paradata seem likely to be of increasing importance.

Asking scalar questions in surveys has always been a significant challenge. The proliferation of the Web during the age of information technology, where how the survey appears on individual computers or other devices differs from what the survey designer sees, means that the challenge is even greater. However, in addition to the challenges, the Web provides a multitude of opportunities for further research on the visual presentation of survey questions, particularly since experiments can be integrated easily and at a relatively low cost. Two specific areas for future exploration are understanding how verbal information and visual design interact to influence respondents and which specific question layouts and wordings are easier for respondents to process. Although much has been learned about these complexities, more remains to be understood in order

for us to achieve accurate measurement of opinions and attitudes that are central to research in sociology and the other social sciences.

## References

Christian, L. M. 2007. "How Mixed-Mode Surveys Are Transforming Social Research: The Influence of Survey Mode on Measurement in Web and Telephone Surveys." PhD dissertation, Washington State University, Pullman, WA.

Christian, Leah M. and Don A. Dillman. 2004. "The Influence of Symbolic and Graphical Language Manipulations on Answers to Paper Self-Administered Questionnaires." *Public Opinion Quarterly* 68:57-80.

Christian, Leah Melani, Don A. Dillman, and Jolene D. Smyth. 2007a. "The Effects of Mode and Format on Answers to Scalar Questions in Telephone and Web Surveys." Pp. 250-75 in *Advances in Telephone Survey Methodology*, edited by James M. Lepkowski, Clyde Tucker, J. Michael Brick, Edith de Leeuw, Lilli Japec, Paul. J. Lavrakas, Michael W. Link, and Roberta L. Sangster. New York: John Wiley.

———. 2007b. "Helping Respondents Get It Right the First Time: The Influence of Words, Symbols, and Graphics in Web Surveys." *Public Opinion Quarterly* 71:113-25.

Couper, Mick P. 2005. "Technology Trends in Survey Data Collection." *Social Science Computer Review* 23:486-501.

Couper, Mick P., Roger Tourangeau, Fred G. Conrad, and Scott D. Crawford. 2004. "What They See Is What We Get: Response Options for Web Surveys." *Social Science Computer Review* 22:111-27.

Couper, Mick P., Roger Tourangeau, Fred G. Conrad, and Eleanor Singer. 2006. "Evaluating the Effectiveness of Visual Analog Scales: A Web Experiment." *Social Science Computer Review* 24:227-45.

Dillman, Don A., Tommy L. Brown, John Carlson, Edwin H. Carpenter, Frederick O. Lorenz, Robert Mason, John Saltiel, and Roberta L. Sangster. 1995. "Effects of Category Order on Answers to Mail and Telephone Surveys." *Rural Sociology* 60:674-87.

Dillman, Don A., Glenn Phelps, Robert Tortora, Karen Swift, Julie Kohrell, Jodi Berck, and Benjamin L. Messer. Forthcoming. "Response Rate and Measurement Differences in Mixed Mode Surveys Using Mail, Telephone, Interactive Voice Response (IVR) and the Internet." *Social Science Research.*

Dillman, Don A., Roberta L. Sangster, John Tarnai, and Todd Rockwood. 1996. "Understanding Differences in People's Answers to Telephone and Mail Surveys." Pp. 45-62 in *Current Issues in Survey Research*, edited by Marc T. Braverman and Jana Kay Slater. San Francisco: Jossey-Bass.

Dillman, Don A., Jolene D. Smyth, and Leah Melani Christian. 2009. *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method.* 3rd ed. New York: John Wiley.

Fuchs, Marek. 2008. "Gender of Interviewer Effects in Video-Enhanced Web Surveys: Results From a Randomized Field-Experiment." Presented at the Annual General Online Research Conference, Hamburg, Germany.

Heerwegh, Dirk. 2003. "Explaining Response Latency and Changing Answers Using Client Side Paradata From a Web Survey." *Social Science Computer Review* 21:360-73.

Heerwegh, Dirk and G. Loosveldt. 2002. "An Evaluation of the Effects of Response Formats on Data Quality in Web Surveys." *Social Science Computer Review* 20:469-82.

Israel, Glen D. and C. L. Taylor. 1990. "Can Response Order Bias Evaluations?" *Evaluation and Program Planning* 13:1-7.

Jenkins, Cleo and Don A. Dillman. 1997. "Towards a Theory of Self-Administered Questionnaire Design." Pp. 165-96 in *Survey Measurement and Process Quality*, edited by Lars E. Lyberg, Paul Biemer, Martin Collins, Edith D. de Leeuw, Cathryn Dippo, Norbert Schwarz, and Dennis Trewin. New York: Wiley InterScience.

Johnson, Martin. 2004. "Timepieces: Components of Survey Question Response Latencies." *Political Psychology* 25:679-702.

Krosnick, Jon A. and Duane F. Alwin. 1987. "An Evaluation of a Cognitive Theory of Response Order Effects in Survey Measurement." *Public Opinion Quarterly* 51:201-19.

Krosnick, Jon A. and L. R. Fabrigar. 1997. "Designing Rating Scales for Effective Measurement in Surveys." Pp. 141-64 in *Survey Measurement and Process Quality*, edited by Lars E. Lyberg, Paul Biemer, Martin Collins, Edith D. de Leeuw, Cathryn Dippo, Norbert Schwarz, and Dennis Trewin. New York: Wiley InterScience.

Krysan, Maria and Mick P. Couper. 2003. "Race in the Live and Virtual Interviewer: Racial Deference, Social Desirability, and Activation Effects in Attitude Surveys." *Social Psychology Quarterly* 66:364-83.

Mahon-Haft, Taj and Don A. Dillman. 2007. "Isolating Primacy-Inducing Conditions in Web Surveys. 2007." Presented at the American Association for Public Opinion Research Annual Conference, Anaheim, CA.

Redline, Cleo and Don A. Dillman. 2002. "The Influence of Alternative Visual Designs on Respondents' Performance With Branching Instructions in Self-Administered Questionnaires." Pp. 179-93 in *Survey Nonresponse*, edited by Robert M. Groves, Don A. Dillman, John L. Eltinge, and Roderick J. A. Little. New York: John Wiley.

Redline, Cleo, Don A. Dillman, Aref N. Dajani, and Mary Ann Scaggs. 2003. "Improving Navigational Performance in U.S. Census 2000 by Altering the Visually Administered Languages of Branching Instructions." *Journal of Official Statistics* 19:403-20.

Schober, Michael F. and Fred G. Conrad. 2007. "Dialogue Capability and Perceptual Realism in Survey Interviewing Agents." Presented at the annual meeting of the American Association for Public Opinion Research, Montreal, Canada.

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

Schwarz, Norbert, C. E. Grayson, and B. Knäuper. 1998. "Formal Features of Rating Scales and the Interpretation of Question Meaning." *International Journal of Public Opinion Research* 10:177-83.

Schwarz, Norbert, B. Knäuper, H. J. Hippler, E. Noelle-Neumann, and F. Clark. 1991. "Rating Scales: Numeric Values May Change the Meaning of Scale Labels." *Public Opinion Quarterly* 55:570-82.

Smith, Tom W. 1995. "Little Things Matter: A Sample of How Differences in Questionnaire Format Can Affect Survey Responses." Presented at the American Association for Public Opinion Research Annual Conference, Ft. Lauderdale, FL.

Smyth, Jolene D., Don A. Dillman, Leah Melani Christian, and Michael J. Stern. 2006. "Comparing Check-All and Forced-Choice Question Formats in Web Surveys." *Public Opinion Quarterly* 70:66-77.

Stern, Michael J. 2008. "The Use of Client Side Paradata in Analyzing the Effects of Visual Layout on Changing Responses in Web Surveys." *Field Methods* 20:377-98.

Stern, Michael J., Don A. Dillman, and Jolene D. Smyth. 2007. "Visual Design, Order Effects, and Respondent Characteristics in a Self-Administered Survey." *Survey Research Methods* 1:121-38.

Thomas, Randall K. and Mick P. Couper. 2007. "A Comparison of Visual Analog and Graphic Ratings Scales." Presented at the General Online Research Conference, Leipzig, Germany.

Toepoel, Vera. 2008. A Closer Look at Web Questionnaire Design. The Netherlands Dissertation Series, Center for Economic Research, Tilburg University.

Tourangeau, Roger, Mick P. Couper, and Fred G. Conrad. 2004. "Spacing, Position, and Order. Interpretive Heuristics for Visual Features of Survey Questions." *Public Opinion Quarterly* 68:368-93.

———. 2007. "Color, Labels, and Interpretive Heuristics for Response Scales." *Public Opinion Quarterly* 71:91-112.

Ware, Colin. 2004. *Information Visualization: Perception for Design*. San Francisco: Morgan Kaufmann.

**Leah Melani Christian** is a research associate at the Pew Research Center for the People & the Press in Washington, D.C. Her research interests include public opinion and political sociology, globalization and technology, and research methods. Her specialty areas within survey methodology are web and mixed-mode surveys, questionnaire design, visual design theory, and cell phone interviewing.

**Nicholas L. Parsons** is an assistant professor of Sociology at Eastern Connecticut State University. Currently, his primary research interest is in the sociology of drug use, particularly the history and cultural evolution of methamphetamine in the United States. His other scholarly interests include criminology, sociology of sport, collective memory, research methods, and sociology of culture.

**Don A. Dillman** is Regents Professor and the Thomas S. Foley Distinguished Professor of Government and Public Policy at Washington State University. Since 1993 he has maintained an extensive research program on how visual layout and design affects answers to web and mail surveys. His current research emphasizes finding improved ways of using the Web in mixed-mode survey designs.